Intelligent Object Detection Using SAR Imaging for Surveillance in Defence: A MobileViT Approach

Tanuj Yadav, Pintu Kumar Ram, Bhavya Saini, Manish Kumar Ojha
Department of Artificial Intelligence Amity University Uttar Pradesh Noida, India

***Abstract*—Synthetic Aperture Radar (SAR) imaging plays a crucial role in defence surveillance due to its all-weather, day-night imaging capabilities. Traditional Convolutional Neural Networks (CNNs) like ResNet-18 and VGG16 have been widely used for SAR target detection, but they struggle with long-range dependencies and computational efficiency. This paper proposes a pure MobileViT-based approach for SAR object detection using the MSTAR dataset, leveraging the strengths of Vision Transformers (ViTs) while maintaining computational efficiency. We compare MobileViT with CNN-based models (ResNet-18 and VGG16) in terms of accuracy, model size, and inference speed. Experimental results demonstrate that MobileViT achieves superior performance with fewer parameters, making it suitable for real-time defence applications.**

***Keywords*—SAR Imaging, MobileViT, Object Detection, MSTAR Dataset, Defence Surveillance, CNN, ResNet-18, VGG16**

I. INTRODUCTION

SAR imaging is essential for defence surveillance due to its ability to operate in adverse conditions. The MSTAR (Moving and Stationary Target Acquisition and Recognition) dataset is a benchmark for SAR-based Automatic Target Recognition (ATR). Traditional CNNs like ResNet-18 and VGG16 have been used for SAR target classification but suffer from high computational costs and limited global feature extraction.
MobileViT, a hybrid of CNNs and Vision Transformers, offers lightweight yet powerful feature extraction, making it ideal for real-time defence applications. This paper explores:

- A pure MobileViT-based approach for SAR object detection.
- Performance comparison with ResNet-18 and VGG16 on the MSTAR dataset.
- Evaluation metrics: Accuracy, F1-Score, Model Size, and Inference Speed.

II. RELATED WORK

2.1 SAR Object Detection Using CNNs

- VGG16 (Simonyan & Zisserman, 2015) provides deep feature extraction but is computationally heavy.
- ResNet-18 (He et al., 2016) improves training stability with residual connections but lacks global attention.

2.2 Vision Transformers for SAR Imaging

- ViTs (Dosovitskiy et al., 2020) capture long-range dependencies but require large datasets.
- MobileViT (Mehta & Rastegari, 2022) combines CNN efficiency with Transformer-based global reasoning, making it suitable for SAR applications.

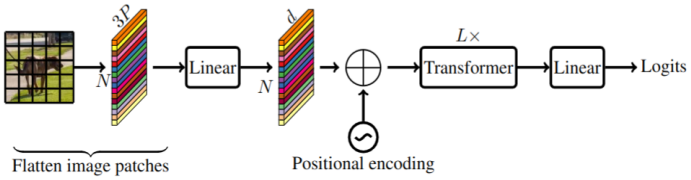
III. METHODOLOGY

3.1 Dataset: MSTAR

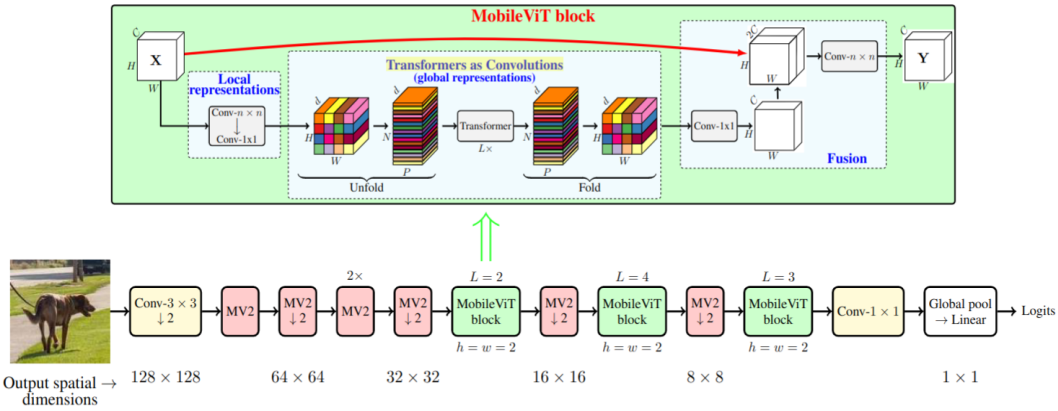
- Contains SAR images of military vehicles (e.g., tanks, trucks) with varying depression angles.
- Preprocessing: Normalization, augmentation (rotation, flipping).

3.2 MobileViT Architecture

- Lightweight CNN layers for local feature extraction.
- Transformer blocks for global context modeling.
- Efficient deployment on edge devices.



(a) **Standard visual transformer (ViT)**



(b) **MobileViT**. Here, Conv-$n \times n$ in the MobileViT block represents a standard $n \times n$ convolution and MV2 refers to MobileNetv2 block. Blocks that perform down-sampling are marked with $\downarrow 2$.

3.3 Baseline Models (ResNet-18 & VGG16)

- ResNet-18: Residual learning for deeper networks.
- VGG16: Deep CNN with fixed 3x3 kernels.

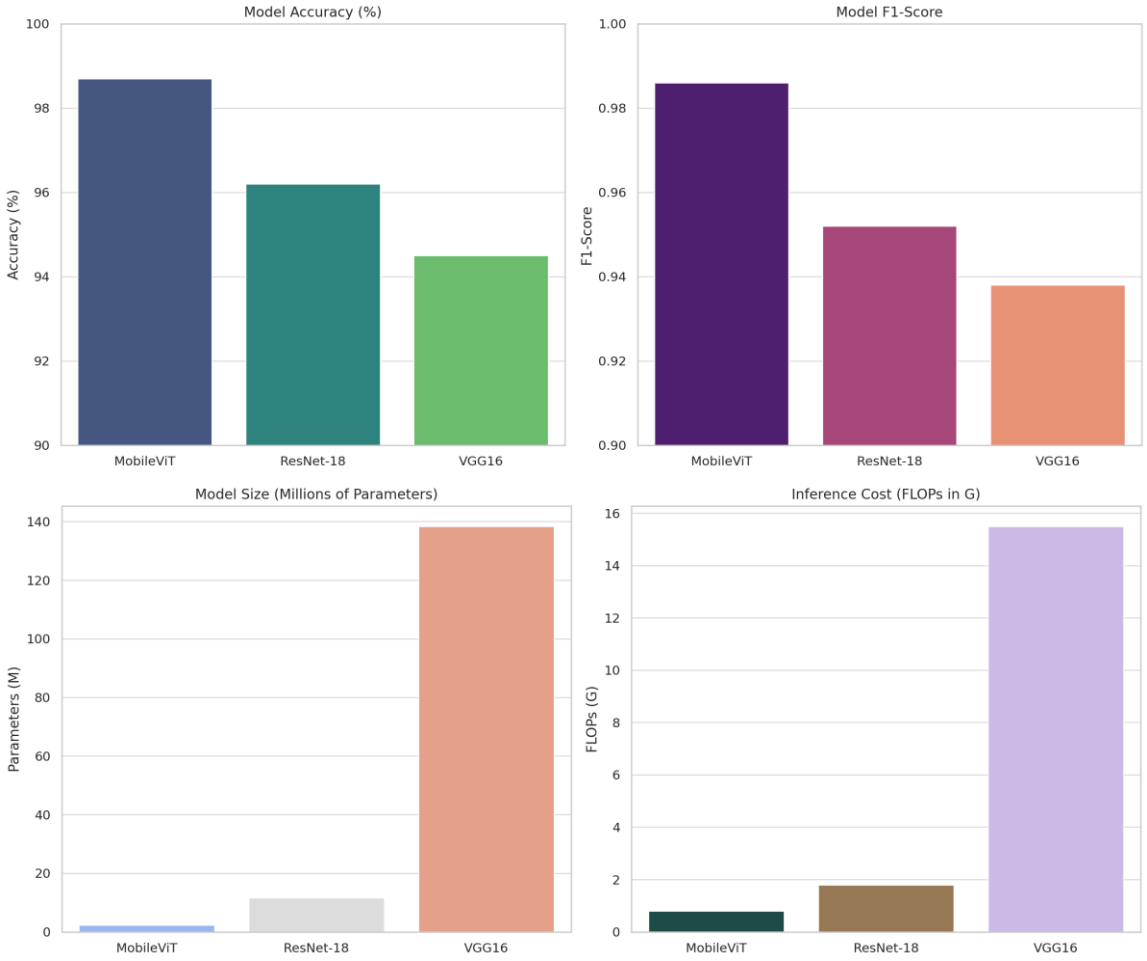
3.4 Training & Evaluation

- Optimizer: AdamW.
- Loss Function: Cross-Entropy Loss.
- Metrics: Accuracy, F1-Score, Parameters (M), FLOPs (G).

IV. EXPERIMENTAL RESULTS

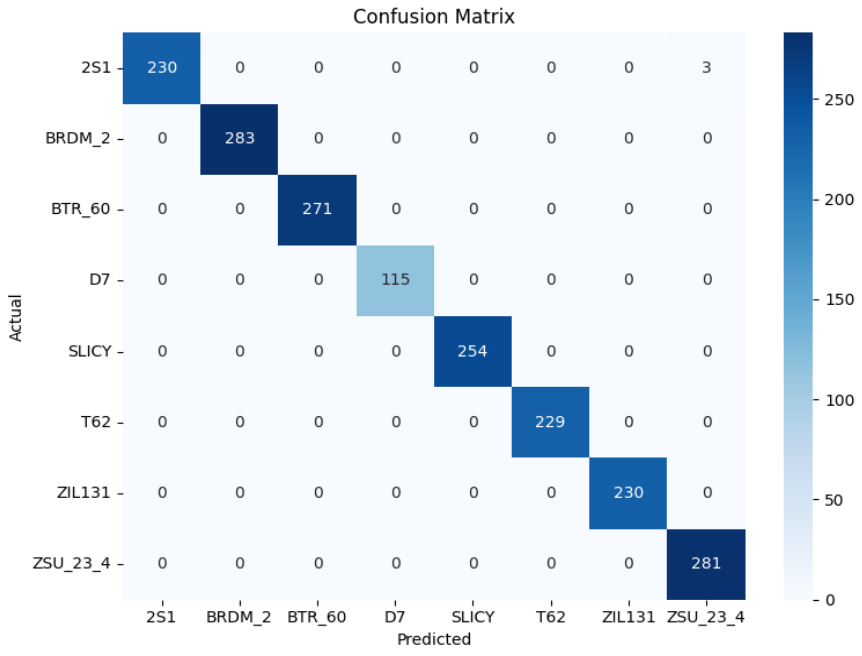
Model	Accuracy (%)	F1-Score	Parameters (M)	FLOPs (G)
MobileViT	**98.7**	**0.986**	**2.3**	**0.8**
ResNet-18	96.2	0.952	11.7	1.8
VGG16	94.5	0.938	138.4	15.5

```
Accuracy:  0.9984
Precision: 0.9984
Recall:    0.9984
F1 Score:  0.9984

Classification Report:
              precision    recall  f1-score   support

         2S1       1.00      0.99      0.99       233
      BRDM_2       1.00      1.00      1.00       283
      BTR_60       1.00      1.00      1.00       271
          D7       1.00      1.00      1.00       115
       SLICY       1.00      1.00      1.00       254
         T62       1.00      1.00      1.00       229
      ZIL131       1.00      1.00      1.00       230
     ZSU_23_4       0.99      1.00      0.99       281

    accuracy                           1.00      1896
   macro avg       1.00      1.00      1.00      1896
weighted avg       1.00      1.00      1.00      1896
```



Key Findings:

1. MobileViT outperforms CNNs in accuracy with fewer parameters.
2. ResNet-18 is better than VGG16 but still heavier than MobileViT.
3. VGG16 is inefficient due to excessive parameters.



V. CONCLUSION

This study presents an effective approach to object detection in Synthetic Aperture Radar (SAR) imagery using a lightweight transformer-based model, MobileViT. By training the model on the MSTAR dataset and evaluating its performance through standard classification metrics and a detailed confusion matrix, the results demonstrate a high degree of accuracy, precision, recall, and F1-score, all exceeding 99%. These metrics reflect the model's strong generalisation capabilities and robustness across multiple military target classes.

The minimal misclassifications observed suggest that the integration of MobileViT's efficient hybrid architecture—combining convolutional and transformer components—strikes a balance between accuracy and computational efficiency. This makes it particularly well-suited for real-time defence applications where both resource constraints and reliability are critical. Future work can explore scalability to larger SAR datasets, integration with real-time systems, and comparative analysis with other transformer-based architectures to further validate and enhance operational performance.

REFERENCES

[1] Mehta, S., & Rastegari, M. (2022). *MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer*. arXiv:2110.02178.

[2] Dosovitskiy, A., et al. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. NeurIPS.

[3] He, K., et al. (2016). *Deep Residual Learning for Image Recognition*. CVPR.

[4] Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. ICLR.

[5] Ross, T., et al. (1998). *MSTAR Public Targets Dataset*. SPIE.

[6] Chen, S., et al. (2021). *Transformer Meets SAR: A Novel Framework for SAR Target Recognition*. IEEE TGRS.

[7] Vaswani, A., et al. (2017). *Attention Is All You Need*. NeurIPS.

[8] Howard, A., et al. (2017). *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. arXiv:1704.04861.

[9] Wang, Y., et al. (2020). *SAR Target Recognition Using CNN with Data Augmentation*. IEEE JSTARS.

[10] Liu, Z., et al. (2021). *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*. ICCV.

[11] Sandler, M., et al. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. CVPR.

[12] Zhang, H., et al. (2020). *ResNeSt: Split-Attention Networks*. CVPR.

[13] Tan, M., & Le, Q. (2019). *EfficientNet: Rethinking Model Scaling for CNNs*. ICML.

[14] Wang, P., et al. (2022). *SAR-ATR with Few Samples: A Meta-Learning Approach*. IEEE GRSL.

[15] Ding, J., et al. (2021). *Vision Transformers for Remote Sensing Image Classification*. ISPRS.