Aerial Landscapes

Shixun Li (z5505146) Xinbo Li (z5496624) Richard Lai (z5620374) Qiyun Li (z5504759) Junle Zhao(z5447039)

Abstract

This project systematically evaluated the performance of six classification models on the SkyView aerial landscape image dataset, which contains 12,000 images from 15 different land categories: agriculture, airport, beach, city, desert, forest, grassland, highway, lake, mountain, parking, port, railway, residential, and river. The models tested included the classic K-Nearest Neighbours (KNN) technique, ResNet-18, a Custom lightweight convolutional neural network (CNN6-GAP), and three popular deep learning architectures: DenseNet-121, Swin Transformer, and ViT-Base. All models were built with the PyTorch framework and fine-tuned via a uniform training pipeline. Modern training methods were used, such as data augmentation, label smoothing, the AdamW optimizer, learning rate schedulers, and automatic mixed precision training. Regarding test results, ViT-Base-224 achieved the highest classification accuracy (99.58%) and used its powerful global modelling capabilities to outperform all other models. The Swin Transformer was close behind with 97.38%, attributed to its modified window-focus mechanism and hierarchical structure. The DenseNet-121 achieved 98.50% accuracy, demonstrating a steady reuse of components, facilitated by its compact modular design. The ResNet-18 achieved 97.68% accuracy and served as a reliable baseline model with a classic residual pattern. A custom-built CNN (CNN6-GAP) achieved a precision of 91.17%, demonstrating good performance and usability in resource-intensive environments. Although the traditional SVM method was less accurate (61.0%), it offers some advantages in terms of the interpretability of the model. The study highlights the trade-off between the classification accuracy and the model complexity, shows the efficiency of modern deep learning approaches for aerial image classification tasks and provides practical insights and technical guidance for the future selection and implementation of the model in real-world applications.

Keywords

SkyView Dataset, Aerial Image Classification, Remote Sensing Images, Convolutional Neural Network (CNN), Vision Transformer, Swin Transformer, ViT, DenseNet, ResNet, Image Recognition

I. INTRODUCTION

Aerial image classification is a key task at remote sensing and computer vision intersection, with wide-ranging applications in real-world scenarios such as urban planning, land-use analysis, environmental monitoring, and disaster response. However, aerial images often exhibit challenges such as varying shooting angles, complex terrain structures, and diverse lighting conditions, significantly hindering image understanding. These complexities limit the accuracy and generalization capability of traditional classification methods.

In recent years, with the rapid development of deep learning, models based on Convolutional Neural Networks (CNNs) and Transformers have achieved breakthroughs in image recognition tasks, demonstrating powerful feature extraction and semantic modelling capabilities. Nevertheless, there remains a trade-off between performance, model complexity, and computational resource consumption. Therefore, selecting the most suitable model for aerial image classification thus remains a subject worthy of systematic investigation.

Toward this objective, we explored and compared six classification models using the SkyView aerial landscape image dataset, including SVM, ResNet-18, DenseNet-121, Swin Transformer, ViT-Base, and a custom lightweight CNN and so on. They have same training pipeline, and their adaptability and deployment value in practical scenarios were analyzed.

II. RELATED WORK

LeNet-5 (LeCun et al., 1998) was the first model to successfully apply Convolutional Neural Networks (CNNs) to document recognition tasks, demonstrating the effectiveness of gradient descent in image classification and laying the development foundation for deep learning in computer vision.

DenseNet (Huang et al., 2017) proposed a dense connectivity mechanism that establishes cross-layer connections within the network, effectively enhancing feature reuse and gradient propagation. This model significantly improves training efficiency and classification performance in deep networks.

Swin Transformer (Liu et al., 2021) innovatively introduced shifted windows and a hierarchical structure into the Vision Transformer architecture. This model enables a balance between local and global modelling while significantly improving computational efficiency, making it particularly suitable for high-resolution images.

Vision Transformer (Dosovitskiy et al., 2021) was the first to apply a pure Transformer architecture to image recognition tasks, surpassing traditional CNNs under large-scale pretraining and marking the beginning of the Vision Transformer era.

ResNet (He et al., 2015) introduced residual connections, effectively addressing the problems of vanishing gradients and performance degradation in deep networks. This model successfully trained very deep neural networks and led to remarkable achievements in standard image recognition tasks.

SVM (Wang and Hu, 2005) played a significant role in early image classification research. Its strong interpretability and generalization ability make it still applicable in low-dimensional feature spaces. However, it doesn't working well for large-scale, high-dimensional image data.

To evaluate the performance of these models in practical applications, this studycompares six classification methods on the SkyView multi-class aerial image dataset: DenseNet-121, Swin Transformer, ViT-Base, ResNet-18, a custom lightweight convolutional model (PlusExtremeCNN), and the traditional Support Vector Machine (SVM). All models were trained and evaluated using a unified training pipeline. Modern training strategies—including label smoothing, the AdamW optimizer, learning rate schedule, and automatic mixed precision—were applied, along with diverse data augmentation techniques, such as random cropping, rotation, and brightness adjustment, to enhance robustness and generalization. Furthermore, the custom PlusExtremeCNN model maintained good recognition performance while consuming minimal computational resources, making it suitable for deployment in resource-constrained environments. The systematic evaluation presented in this study offers valuable insights and technical guidance for model selection and practical engineering in aerial image classification tasks.

III. METHODS

This project first introduced the traditional machine learning method—Support Vector Machine (SVM)—as a baseline for comparative analysis compare to different way to extracting features.We chose the extracting SIFT features and employing a Bag-of-Words (BoW) model to convert images into fixed-length visual word vectors. MiniBatchKMeans clustering was used for vocabulary encoding. Then we used Principal Component Analysis (PCA) to reduce the feature dimensionality to 60 components. While this method offers good interpretability and intuitive modelling, it lacks end-to-end learning capability and struggles with high-dimensional and complex image data, achieving only 61.0% accuracy on the test set. Without PCA, the accuracy drops to 59.9%. This highlights the limitations of traditional methods in remote sensing image recognition tasks.

In addition, Local Binary Pattern (LBP) was evaluated as an feature descriptor. LBP captures local texture by compare neighborhood around every pixel, and record the histograms of grayscale image. We find features by skimage.feature.local_binary_pattern, then we used PAC and standardization same as the SIFT. As a baseline, a linear SVM was used for classification,this method achieved a test accuracy of 60.8%, slightly below the SIFT.

We also try different traditional classifiers, including k-Nearest Neighbors (k-NN) and Random Forest (RF) using the same SIFT-based BoW features. With k=3, the k-NN model achieved a low test accuracy of 41.3%; In the RF, we configured with 100 trees, it achieved 55.7% accuracy. These results indicate that the RF model is better than KNN model on complex patterns, but they still fall short of SVM.

In order to improve feature representation and training efficiency, we designed a custom lightweight convolutional neural network, CNN6-GAP. This network has of six convolutional modules (Conv → BatchNorm → ReLU → MaxPool) for extracting local image features, followed by Global Average Pooling (GAP) to reduce dimensionality and mitigate overfitting. The classifier comprises two fully connected layers (1024 → 512 → number of classes). CNN6-GAP is compact and fast, making it well-suited for deployment on edge devices or in resource-constrained environments. However, it has limited capacity in handling complex textures or fine-grained structures.

Building on CNN6-GAP, we introduced the classic deep residual network ResNet-18 to enhance feature extraction quality. ResNet-18 leverages residual connections to alleviate the vanishing gradient problem in deep networks, providing strong stability and transferability. We used pre-trained weights from ImageNet-1K and fine-tuned the entire network. As a lightweight CNN representative, ResNet-18 outperformed traditional methods significantly, achieving 97.67% accuracy on the test set while maintaining efficiency.

To further enhance model capacity, we adopted the DenseNet-121 architecture. By introducing dense connectivity, this model strengthens feature reuse and gradient flow, effectively reducing information degradation across layers. We utilized the pre-trained weights from torchvision (ImageNet-1K) and fine-tuned the entire model after replacing the classification head. Label smoothing ($\varepsilon = 0.1$) was also applied to improve generalization. DenseNet-121 performed our task excellently, achieving 98.50% validation accuracy. However, it remains limited by the inherent locality of CNNs, restricting its ability to model global semantic relationships.

To address the limitations of convolutional networks in capturing long-range dependencies and global context, we introduced the Swin Transformer. This model integrates local attention and a hierarchical structure through shifted window attention mechanisms, balancing computational efficiency and local feature expressiveness. We loaded the pre-trained Swin-Base model from the timm library (trained on ImageNet-22k) and performed full fine-tuning. Swin Transformer achieved a well-balanced performance in our study, reaching 97.54% validation accuracy with controlled computational cost.

Finally, we employed the Vision Transformer (ViT-Base-Patch16-224-In21k), a model with superior global modelling capabilities. ViT segments images into patch sequences and feeds them into a pure Transformer architecture, leveraging self-attention mechanisms for global semantic modelling. We used pre-trained weights from Hugging Face (trained on ImageNet-21k) and conducted full fine-tuning. ViT exhibited strong representation power for high-resolution images and achieved the highest accuracy in this study (99.58%). However, it also required substantial training data and computational resources, with convergence and robustness requiring close monitoring.

In summary, this project began with traditional methods (SVM) and progressively introduced lightweight CNNs, deeper convolutional networks, residual networks, densely connected networks, and two types of Transformer-based architectures. We established a comprehensive framework for evaluating image classification models, systematically exploring network design, transfer learning, training optimization strategies, and performance assessment. The study provides clear comparative experiments and practical insights for multi-class aerial image classification and reveals the trade-offs between accuracy, efficiency, and deployment feasibility across different model types.

IV. EXPERIMENTS

SkyView Remote Sensing Image Classification Project: Dataset Description and Model Evaluation Strategy.

Dataset Description and Challenges:
This project uses the SkyView aerial landscape image dataset, which includes 12,000 high-resolution images across 15 terrain categories, such as agriculture, airport, beach, city, desert, forest, grassland, highway, lake, mountain, parking, port, railway, residential, and river. We adopt a consistent 80% training, and 20% validation split to ensure fairness and comparability across all models.Dataset download link: https://www.kaggle.com/datasets/ankit1743/skyview-an-aerial-landscape-dataset

Key challenges of the dataset include:

Complex terrain and diverse textures: High intra-class variation and inter-class similarity make classification difficult and require strong discriminative modelling capability.

Inconsistent image scales and resolutions: Images must be resized and standardized before being input to the models.

Severe class imbalance: Some categories have significantly fewer samples, which may lead to biased training.

Training Strategy and Evaluation Metrics
All images are uniformly resized to 224×224 before training and undergo standard augmentation techniques, such as random cropping, horizontal flipping, color jittering, and rotation.

Loss Function: Cross-entropy loss (DenseNet incorporates label smoothing)

Optimizers: Adam / AdamW

Learning Rate Schedulers: Fixed or CosineAnnealingLR

Mixed Precision Training: Enabled (AMP) to reduce memory consumption

Evaluation Metrics: Top-1 Accuracy, Macro-F1 Score, Cross-Entropy Loss

To address class imbalance, we introduce class-specific weights into the cross-entropy loss function. The model checkpoints are saved based on the best validation Macro-F1 score or lowest validation loss.

Robustness Testing and Explainability
To explore the real-world stability and practical usability of the models, we conducted robustness testing on CNN6-GAP and DenseNet-121 under three types of image perturbations and used Grad-CAM to visualize and explain model behavior.

A. Model Explainability (Grad-CAM Heatmaps)

Grad-CAM heatmaps were generated to visualize the model's attention regions. In most classes, the model successfully focused on semantically meaningful regions such as buildings, roads, or treetops, demonstrating its ability to learn discriminative features. Some misclassified samples exhibited semantic confusion, reflecting the challenges of fine-grained recognition in remote sensing imagery.

B. Robustness Testing

To further assess robustness under realistic conditions, we subjected **CNN6-GAP** and **DenseNet-121** to three common types of image perturbations:

1. Gaussian Noise: Simulates signal noise interference during image acquisition or transmission.

2. Gaussian Blur: Simulates detail loss due to defocus or motion blur.

3. Occlusion: Simulates partial obstruction (e.g., shadows or clouds) by randomly overlaying black patches on the image.

V. RESULTS

Model Performance Summary:
SVM: Utilizes SIFT + Bag-of-Words features with a linear SVM classifier, achieving an accuracy of 61.0%. Serves as a traditional baseline method.

CNN6-GAP: A custom lightweight CNN model, achieving 91.17% accuracy. Suitable for edge deployment.

ResNet-18: A classical CNN architecture, achieving 97.67% accuracy. Stable and reliable.

DenseNet-121: Employs densely connected layers, achieving 98.50% accuracy with strong generalization ability.

Swin Transformer: Introduces shifted window attention and a hierarchical structure, reaching 97.54% accuracy.

ViT (Vision Transformer): Demonstrates the strongest global modelling capability, achieving the highest accuracy of 99.58%.

Robustness Analysis

CNN6-GAP:
CNN6-GAP demonstrated strong robustness to occlusion and Gaussian noise, achieving classification accuracies of 89.4% and 69.3%, respectively, under these perturbations. This performance can be attributed to the inclusion of random occlusion, noise, and blur augmentation during training, which enhanced the model's tolerance to such disturbances. However, despite incorporating blur augmentation, the model's accuracy on blurred images still dropped to 36.1%, indicating sensitivity to image clarity. This may be due to the relatively shallow architecture of CNN6-GAP, which heavily relies on local texture features for classification. When blur disrupts these textures, the model's feature extraction capability significantly deteriorates.

DenseNet-121:
This model employed a relatively complete set of data augmentation strategies (e.g., random cropping, flipping, rotation, and color jittering), enhancing its adaptability to spatial and lighting variations. However, it did not incorporate robustness-specific augmentations such as noise, blur, or occlusion during training. Consequently, its performance under Gaussian noise and blur perturbations declined, with accuracies dropping to 56.6% and 61.7%, respectively, reflecting weaker resistance to such distortions. In contrast, the model performed relatively well under occlusion (68.2% accuracy). DenseNet can make reasonable predictions due to its deep structure and dense connections, which help capture global semantic information even when some local textures are obstructed.

VI. CONCLUSION

VII. CONTRIBUTIONS:

This project constructed and evaluated six classification models on the SkyView aerial image dataset, covering a wide spectrum of approaches—from traditional methods such as Support Vector Machines (SVM), K-Nearest Neighbours(KNN) to modern deep Transformer-based architectures. The models include a custom lightweight convolutional neural network (CNN6-GAP), ResNet-18, DenseNet-121, Swin Transformer, and ViT-Base. All models were trained under a unified pipeline to ensure fair comparisons. In addition, we still do the robustness testing (including Gaussian noise, blur, and occlusion) and Grad-CAM interpretability analysis were conducted to comprehensively assess the models’ stability and real-world deployment value. At the end, the CNN6-GAP model maintained competitive accuracy with significantly reduced parameters, making it highly suitable for deployment on edge devices and demonstrating strong practical applicability.

Strengths:

1. Broad coverage of model types, ranging from traditional algorithms to state-of-the-art Transformer architectures.
2. All models were trained with consistent data augmentation and training strategies, ensuring scientific and fair comparison.
3. The inclusion of robustness testing and interpretability analysis enhanced the understanding of model performance under complex real-world scenarios.
4. CNN6-GAP is lightweight and deployment-friendly, achieving a practical balance between efficiency and performance.
5. The experiments were stable, with no signs of overfitting, and demonstrated strong convergence and generalization capabilities.

Limitations:

1. The evaluation was conducted on a single dataset (SkyView), so the models' generalization abilities have yet to be validated on other remote sensing datasets.
2. Although Transformer-based models like ViT showed excellent performance, their high computational requirements limit their applicability on low-power platforms.
3. Only certain models (such as CNN6-GAP) incorporated robustness-oriented augmentation techniques, leading to performance fluctuations in others (such as DenseNet-121) under blur and noise conditions.
4. No model fusion experiments were performed, leaving the potential of collaborative optimization among models unexplored.

Future Improvements:

1. Introduce more diverse image degradation simulations (e.g., blur, compression, brightness variation) to enhance robustness to low-quality inputs.
2. Explore lighter Transformer variants (such as Tiny-ViT and MobileViT) to support deployment in embedded or resource-constrained environments.
3. Apply more advanced class imbalance learning techniques (such as focal loss and SMOTE) to improve recognition performance on minority classes.
4. Investigate model fusion strategies (e.g., hybrid CNN and Transformer architectures) to further enhance classification performance.

References

[1] Gradient-based learning applied to document recognition - IEEE Journals & Magazine. [online] Available at: https://ieeexplore.ieee.org/document/726791.

[2] Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K.Q. (2017). Densely Connected Convolutional Networks. [online] openaccess.thecvf.com. Available at: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html.

[3] Liu, Z., Lin, Y., Cao, Y., Hu, H., Yixuan, Zhang, Z., Lin, S. and Guo, B. (n.d.). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. [online] Available at: https://openaccess.thecvf.com/content/ICCV2021/papers/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.pdf.

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs]. [online] Available at: https://arxiv.org/abs/2010.11929v2.

[5] He, K., Zhang, X., Ren, S. and Sun, J. (2015). Deep Residual Learning for Image Recognition. [online] arXiv.org. Available at: https://arxiv.org/abs/1512.03385.

[6] Wang, H. and Hu, D. (2005). Comparison of SVM and LS-SVM for Regression. [online] IEEE Xplore. doi: https://doi.org/10.1109/ICNNB.2005.1614615.