GROUP 38: ROBUSTLY SOLVING AERIAL SCENE CLASSIFICATION

Amna El-Mustafa aelmusta@asu.edu
Amandeep Kaur akaur64@asu.edu
Dan Gibson dggibso1@asu.edu

ABSTRACT

Aerial scene classification is a crucial task in remote sensing, with applications spanning disaster management, urban planning, and environmental monitoring. This project evaluates three distinct families of machine learning models—Convolutional Neural Networks (CNNs), Bayesian Neural Networks (BNNs), and Vision Transformers with Masked Auto-Encoding(MAE ViTs)—on the AID dataset, a benchmark for high-resolution aerial imagery classification. The study emphasizes a comprehensive analysis of performance metrics, uncertainty estimation, and interpretability using Grad-CAM visualizations and confusion matrices. MAE ViTs demonstrated superior classification accuracy (96.9%), leveraging their global contextual understanding through self-attention mechanisms. CNNs achieved competitive results (94.4%) with efficient feature extraction but lacked uncertainty quantification. BNNs, despite lower accuracy (84.13%), provided robust estimates of epistemic and aleatoric uncertainty, making them useful for applications where model reliability is critical.

1 Introduction

Analyzing and categorizing aerial imagery has become a critical capability in remote sensing, offering invaluable support across emergency preparedness, urban development, and environmental conservation. This task involves distinguishing between diverse scene types, which is challenging due to high intra-class variability, complex visual features, and the need for reliable predictions. In this project, we explore and evaluate the performance of three model families—CNNs, BNNs, and (MAEs)—on the AID dataset for aerial scene classification.

Deep learning models, such as CNNs, MAEs [5] have demonstrated exceptional performance in extracting spatial features for classification. However, these models often lack the capability to quantify uncertainty, which is essential in high-stakes scenarios where misclassification carries significant consequences.

BNNs extend traditional deep learning by introducing stochasticity in model parameters, enabling the estimation of both epistemic (model-related) and aleatoric (data-related) uncertainties. By leveraging probabilistic modeling, BNNs provide a more robust framework for applications requiring high confidence in predictions.

In this task, MAE outperformed all other models with accuracy of 96.9%. CNN achieved comparable performance of 94.4% using EffecientNet, while BNN achieved 84.13%, it provided additional uncertainty metrics, assessing the model confidence in its predictions. We also provide analysis for failure cases of the model as well as analyzing where does the model look using some examples and GradCam method [10]. A particular emphasis is placed on analyzing the confusion matrices to understand model reliability and class-specific performance.

1.1 Related work

Aerial scene classification has been studied in the past. For instance, [2] proposes a multiple-instance densely-connected ConvNet (MIDC-Net), which frames aerial scene classification as a multiple-instance learning problem, allowing for deeper exploration of local semantics. Meanwhile, [1] introduces a layer selection strategy called ReLU-Based Feature Fusion (RBFF), which extracts feature maps from MobileNetV2, a CNN-based single-object image classifier, and constructs a model specifically for aerial scene classification. In [3], the authors present the local semantic enhanced ConvNet (LSE-Net), which includes a context-enhanced convolutional feature extractor, a local semantic perception

module, and a classification layer. Further, [14] conducts an empirical study on remote sensing pretraining (RSP) with aerial images, demonstrating how RSP significantly enhances performance in scene recognition and interpretation of remote sensing-specific semantics, like "Bridge" and "Airplane". [15] introduces a novel rotated, varied-size window attention mechanism that replaces the traditional full attention in transformers. This approach effectively reduces computational and memory costs while improving object representation by capturing richer contextual information through diverse windows. All these methods uses AID dataset since it is the largest, more diverse and has highest resolution compared to other datasets as shown in table 1.

Table 1: Comparison of Aerial Scene Classification Datasets [16], [4], [11]

Dataset	**Classes**	**Total Images**	**Resolution**
UC Merced Land Use	21	2,100	256x256
NWPU-RESISC45	45	31,500	256x256
AID	**30**	**10,000**	**600x600**

Our work extends prior studies by incorporating models that enhance robustness, efficiency, and adaptability in remote sensing. First, by adding BNNs to the mix, our work builds on traditional CNN approaches to introduce uncertainty estimation, crucial for high-stakes applications in remote sensing where confidence in classification matters. Additionally, using MAEs enables a self-supervised learning approach, reducing reliance on labeled data and improving generalization for scene classification. This comparison provides a broader perspective on the strengths and trade-offs between standard CNNs, advanced MAEs, and probabilistic BNNs, advancing aerial scene classification methods and understanding in ways that could benefit remote sensing applications.

2 Project Description

We aim to compare three families of machine learning models on the problem of aerial scene classification. We work with BNNs, different CNNs, shown ResNet18 architecture (because if its simplicity) in figure 1 and MAE ViTs in figure 2. We divided the dataset into train, validation and test sets (70%, 15% , 15%). All results shown are for test set.

1. Convolutional Neural Networks (CNNs): CNNs are widely used in image classification tasks due to their ability to efficiently extract spatial features through convolutional operations. In this project, we evaluated several CNN architectures, including ResNet [6], ConvexNet [8], HrNet [12], SeresNet [7] and EfficientNet [13], which are known for their balance between performance and computational efficiency.

   We also did different data augmentation techniques that varies from resizing to flipping vertically and horizontally to random rotations and experimented with and without augmentation. Besides, we experimented with label smoothing loss using a smoothing factor of 0.1 to prevent models from becoming overly confident in their predictions, thereby improving generalization and reducing overfitting. For hyperparameters, ImageNet weights were used to initialize the models and learning rate of 0.001 is used. Adam is used for optimization and the models are trained for 50 epochs.

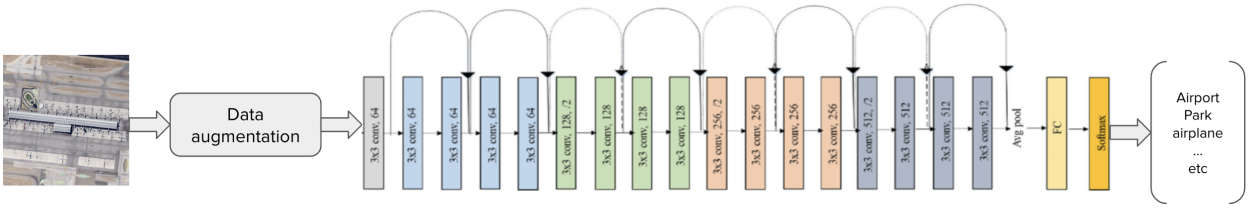

Figure 1: Architectural diagram and Procedure of ResNet18

2. Bayesian Neural Networks (BNNs): BNNs extend traditional neural networks by modeling weights and biases as probability distributions instead of fixed-point estimates. This enables BNNs to estimate both epistemic (model-related) and aleatoric (data-related) uncertainties, making them particularly suitable for scenarios where confidence in predictions is critical. The torchbnn [9] library was used to implement BNNs in this project.
3. Masked Auto-Encoders with Vision Transformers (MAE ViTs: Vision Transformers (ViTs) are a cutting-edge model family that employs transformer-based architectures for image classification. Unlike CNNs, ViTs

use self-attention mechanisms to process entire images, capturing global contextual information effectively. Masked Auto-Encoders (MAEs) [5] are pretraining frameworks for ViTs, where parts of the input image are masked and reconstructed, allowing the model to learn robust representations. Pretrained weights from HuggingFace were leveraged to fine-tune these models for the AID dataset.

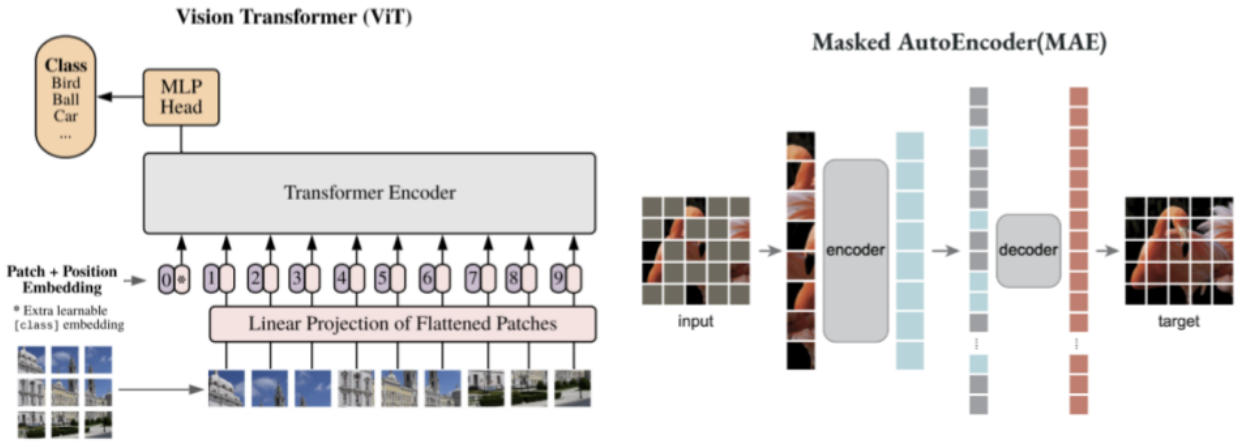

Figure 2: Left: Architectural diagram of ViT | Right: MAE Procedure

3 Achievements

In a semester's time, we were able to understand, code, and compare several models in each of the above families. We list all the tested models per families and highlight the per-family best performance. CNNs model comparisons shown in table 2. MAE ViT model comparisons are shown in table 3

Moreover, the standard metrics, including accuracy, precision, recall, and F1-score, were computed for the best-performing models in each family, shown in table 4.

Table 2: CNN Models Comparison

CNN model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
ResNet18	89.60	89.74	91.02	89.60
ResNet50	92.13	92.15	92.54	92.13
EffecientNet b3	**94.40**	**94.41**	**94.72**	**94.40**
ResNet50- Augmentation	85.93	85.79	86.43	85.93
EffecientNet b3	91.07	91.06	91.39	91.07
ConvexNet	91.47	91.47	91.82	91.47
SeResNet	91.67	91.59	91.93	91.67
HrNet	87.00	87.05	88.13	87.00
ConvexNet- Label smoothing	90.47	90.38	90.54	90.47

Table 3: Result comparison across ViT experiments

Pre-trained weights	No of layers	Hidden size	No of params	Val accuracy
WinKawaks/vit-tiny-patch16-224	12	128	15M	91%
fb/vit-mae-base	12	768	86M	93.2%
facebook/vit-mae-large	**24**	**1024**	**307M**	**96.9%**

Table 4: Standard Metrics for Best-Performing Models

Model Family	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNNs (EfficientNet)	94.40	94.72	94.40	94.41
BNNs	84.13	84.00	84.00	84.00
MAE ViTs	**96.90**	**97.00**	**97.00**	**97.00**

4 Results

We showcase our best results per family in table 4. We also look at per-model confusion matrices to understand the per-class performances and biases. Shown in figures 3, 4 and 8 .The confusion matrix is a very convenient tool to peek into the model's understanding of these classes. Looking at the confusion matrix for the MAE ViT(best model), we notice a few interesting things:

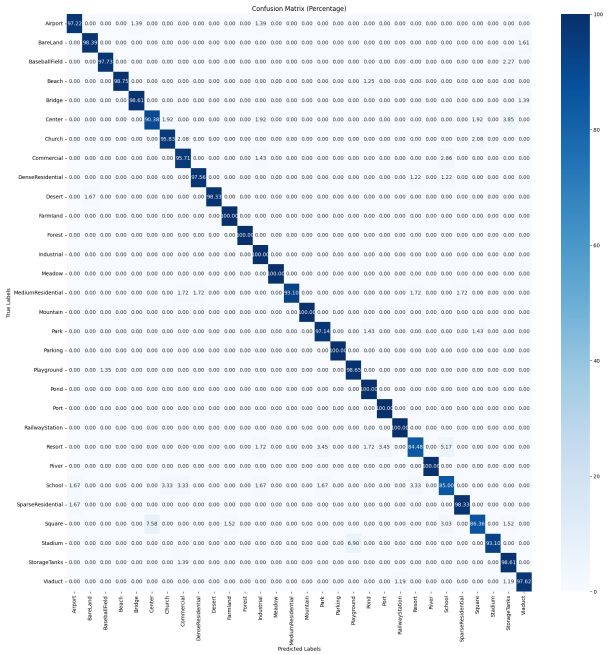

Figure 3: Visualizing confusion matrix for MAE ViT

1. The classes that achieve 100% accuracy are Farmland, Forest, Industrial, Meadow, Mountain, Parking, Pond, Port, Railway Station and River. This is not surprising as we would accept these particular classes to have very strong and distinctive features in overhead imagery. It is interesting to notice that majority of these classes map to natural i.e. not man-made structures.
2. The worst 3 classes i.e. accuracy < 90% are Resort, School and Square which do not have well-defined and distinctive features.
   (a) Resort is mostly confused for School, Port and Park.
   (b) School is most confused for Resort, Church and Commercial
   (c) Square is confused for Center and School. We visualize some of these confusing classes in Figures 5, 6 and 7
3. One can understand that the definitions of some classes can be blurry for example Medium_Residential which sometimes gets confused for Dense_Residential or Sparse_Residential.

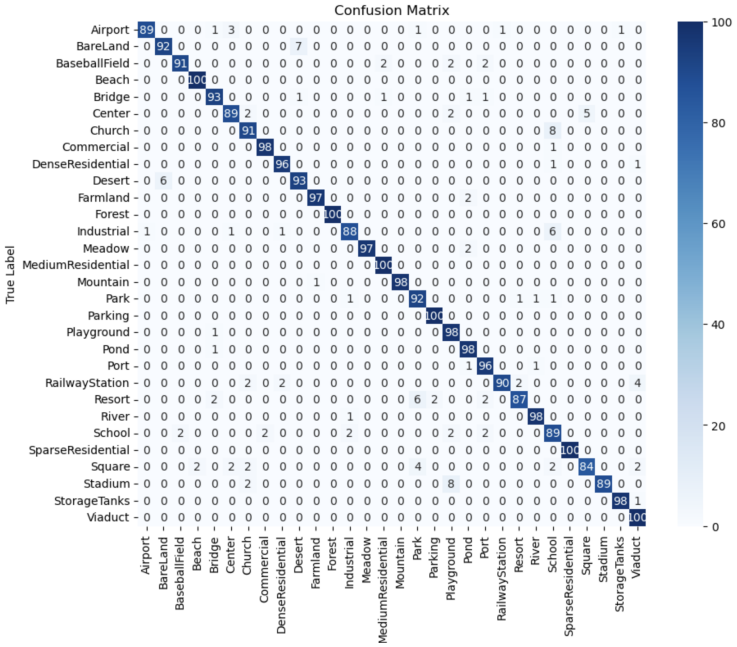

Figure 4: Visualizing confusion matrix for CNN

Figure 5: Visualizing samples misclassified by ViT as Resort along with their predicted labels

To interpret the decision-making process of the models, we generated Grad-CAM visualizations for selected test images. These visualizations highlight the regions of the image that the model focuses on to make its predictions. Shown in figure 9, examples of Grad-CAM outputs for each model. Each row corresponds to a specific model, with three representative test images analyzed. The Grad-CAM visualizations reveal distinct attention patterns for the CNN, BNN, and ViT models. CNNs, known for their proficiency in capturing local features, exhibit a strong focus on edges and textures within the scene. In contrast, BNNs, due to their binary nature, produce noisier and less focused visualizations, indicating potential limitations in precise feature representation. ViTs, on the other hand, demonstrate a more global attention pattern, highlighting multiple regions of the image, showcasing their ability to capture long-range dependencies.

Figure 6: Visualizing samples misclassified by ViT as School along with their predicted labels

Figure 7: Visualizing samples misclassified by ViT as Square along with their predicted labels

4.1 Analysis

Looking at table 4, the MAE ViTs achieved the highest accuracy and F1-scores, demonstrating their effectiveness in capturing global features. CNNs showed competitive performance, particularly with the EfficientNet model. BNNs, while having lower accuracy compared to CNNs and ViTs, provided significant insights into uncertainty, which is crucial in high-stakes decision-making scenarios. The uncertainty metrics highlight the ability of BNNs to differentiate between confident and uncertain predictions, enabling a deeper understanding of the reliability of the model. Table 5 summarizes the average uncertainty values for correct and incorrect predictions, as well as predictive entropy.

For CNNs we can see that EffecientNet b3 without augmentation achieved the best results. It is clear that scaling up the model doesn't necessarily increase the performance for our dataset, as the EffecientNet has one of the lowest number of parameters. We also notice that augmentation caused the performance to decrease for both ResNet50 and EffecientNet. This is a standard problem in remote sensing and satellite imagery where augmentation often distorts the image and introduce noise. Another reason would be that we downscaled the image during augmentation which may result in loss of information and representational power of the model. Label smoothing results in slightly lower performance, which may happen because it reduces the emphasis on the correct class during training, making the model less confident in its predictions. Also another reason would be that we didn't finetune the smoothing factor enough to result in higher performance.

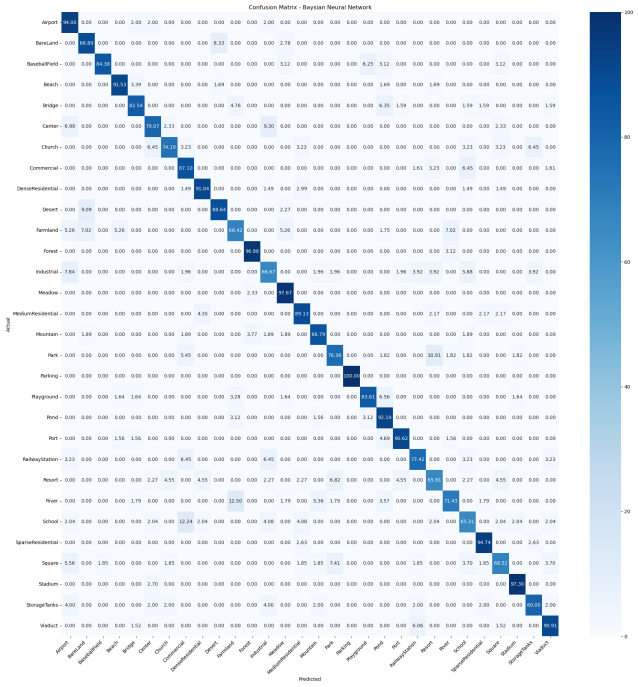

Figure 8: Visualizing confusion matrix for the BNN

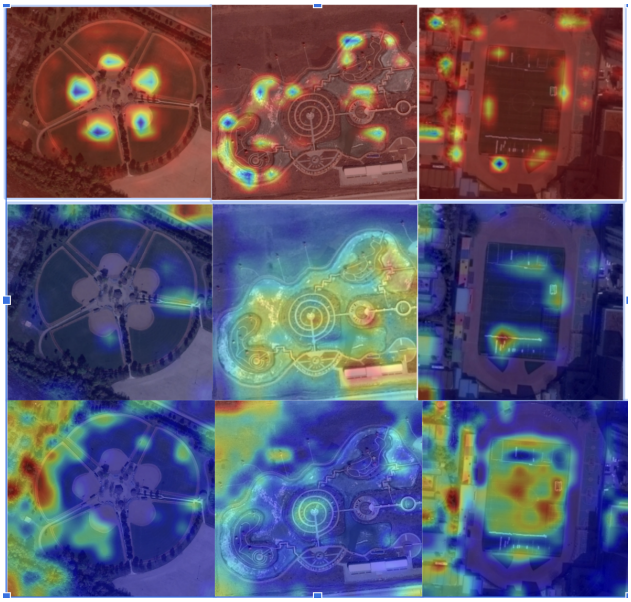

Figure 9: Grad-CAM visualizations for CNN(top) BNN(middle) and VIT(bottom). Labels(Left to right): Baseball Field, Park, Playground

For BNNs, uncertainty helps us understand how confident the model is in its predictions. There are two main types of uncertainty:

- **Epistemic Uncertainty**: This reflects what the model doesn't know due to limited or incomplete training data. For example, if a model hasn't seen many images of a specific scene type, it may be less confident in its classification. Epistemic uncertainty can often be reduced by providing more diverse training data or improving the model.

- **Aleatoric Uncertainty**: This reflects noise or ambiguity in the data itself. For instance, a blurry or low-resolution image can make classification difficult, regardless of how well-trained the model is. This type of uncertainty cannot be reduced because it is inherent to the data.

By interpreting these uncertainties, we can better understand when and why a model might make mistakes. For instance, the higher epistemic and aleatoric uncertainty values for incorrect predictions in Table 5 indicate that the BNN is effectively identifying cases where its predictions are less reliable. This aligns with the trends observed in the confusion matrix. The higher average uncertainty for incorrect predictions (epistemic: 0.0064, aleatoric: 1.2405) compared to correct predictions (epistemic: 0.0029, aleatoric: 0.4949) demonstrates the model's ability to signal a lack of confidence in challenging cases.

Table 5: BNN Uncertainty Metrics

Metric	Correct Predictions	Incorrect Predictions
Epistemic Uncertainty	0.0029	0.0064
Aleatoric Uncertainty	0.4949	1.2405
Predictive Entropy	0.7084	1.7612

5 Member contributions

All team members came together to brainstorm and finalize the project topic, "<Robustly solving aerial scene classification>", perform the literature survey, find the dataset and finalize the model families. Since the plan was code-heavy and involved working with several models, we divided the modeling work as follows:

- Amna handled the understanding, coding and experimenting with CNNs.
- Dan understood and coded up BNNs.
- Aman worked with pretrained weights for ViTs.

After this, we again came together to finalize the methods of comparison and analyze the results of our models.

The presentations and reports were also a result of combined efforts.

6 Conclusion

In this project, we explored and evaluated three model families—Convolutional Neural Networks (CNNs), Bayesian Neural Networks (BNNs), and Masked Auto-Encoders with Vision Transformers (MAE ViTs)—for the challenging task of aerial scene classification using the AID dataset. Each model family brought unique strengths and limitations, highlighting the trade-offs between performance and interpretability.

The MAE ViTs emerged as the best-performing models, achieving the highest accuracy (96.9%) and F1-scores due to their ability to capture global context using transformer-based architectures. CNNs, exemplified by EfficientNet, demonstrated competitive accuracy (94.4%) with efficient feature extraction, although they lack the ability to quantify uncertainty. BNNs, despite their lower accuracy (84.13%), provided robust uncertainty estimates, making them suitable for applications requiring high confidence and reliability in predictions.

Uncertainty analysis revealed that BNNs effectively signal challenging cases through higher epistemic and aleatoric uncertainty values for misclassified samples. This capability underscores their utility in critical applications, such as disaster management and urban planning, where model interpretability and confidence are crucial.

The analysis of confusion matrices provided additional insights into class-specific performance. High-performing classes, such as *Forest* and *Parking*, benefited from distinctive visual features, while low-performing classes, like *Resort* and *Square*, were impacted by overlapping or ambiguous features. Grad-CAM visualizations further illuminated the decision-making process of each model, revealing that MAE ViTs leverage global attention, CNNs focus on localized features, and BNNs demonstrate diffuse attention patterns consistent with their probabilistic nature.

Overall, this project highlights the importance of selecting models that balance accuracy, interpretability, and uncertainty estimation based on application requirements. Future work could explore hybrid models that integrate the uncertainty estimation capabilities of BNNs with the performance of MAE ViTs, as well as strategies to address misclassification in visually similar classes through enhanced feature engineering or data augmentation for remote sensing.

References

[1] Md Adnan Arefeen, Sumaiya Tabassum Nimi, Md Yusuf Sarwar Uddin, and Zhu Li. A lightweight relu-based feature fusion for aerial scene classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3857–3861. IEEE, 2021.

[2] Qi Bi, Kun Qin, Zhili Li, Han Zhang, Kai Xu, and Gui-Song Xia. A multiple-instance densely-connected convnet for aerial scene classification. *IEEE Transactions on Image Processing*, 29:4911–4926, 2020.

[3] Qi Bi, Kun Qin, Han Zhang, and Gui-Song Xia. Local semantic enhanced convnet for aerial scene recognition. *IEEE Transactions on Image Processing*, 30:6498–6511, 2021.

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[8] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. NeurIPS 2019, 33rd Conference on Neural Information Processing Systems, Vancouver, Canada.

[10] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[11] Gaofeng Shao, Junjun Jiang, Xiaoyan Dai, Song Yang, and Peijun Du. Aid: A benchmark dataset for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

[12] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[13] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[14] 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*, 61:1–20, 2022.

[15] 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.

[16] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010.