Vision Transformers for Robust Analysis of Satellite Imagery

Adrian Gamarra
Department of Computer Science
Stanford University
a1090588@stanford.edu

Nathan Kim
Department of Computer Science
Stanford University
nathangk@stanford.edu

Mai Hoang
Department of Computer Science
Stanford University
maihoang@stanford.edu

Abstract

Vision Transformers (ViTs) have shown promise in outperforming traditional Convolutional Neural Networks (CNNs) in image classification [4]. This paper investigates ViTs' performance in land use classification from satellite imagery, focusing on the WILDS Out-of-Distribution (OOD) benchmark [8]. We primarily explore the to Functional Map of the World (FMoW) split from the benchmark as ViT work with images. Our experiments show that ViTs do not outperform the CNN baseline in the context of OOD data. Various strategies to enhance ViT performance were tested, including freezing layers, adjusting learning rates, integrating metadata, and adding dropout, with only minor or no improvements. Our findings suggest that while ViTs offer potential for OOD detection [5], their effectiveness is limited without significant architectural modifications, additional data, or increased computation.

1. Introduction

The Transformer-based model architecture, initially designed for natural language processing (NLP) tasks, has recently been adapted for image classification through Vision Transformers (ViTs). Unlike traditional convolutional neural networks (CNNs), ViTs replace convolutions with self-attention mechanisms. This architectural shift allows ViTs to leverage the computational efficiency and scalability of Transformers, enabling them to outperform state-of-the-art CNNs, especially when pre-trained on extensive datasets. This progress has significantly boosted the adoption of Transformer models in the field of computer vision.

This project aims to use ViTs to improve the robustness of models tasked with classifying land use from satellite imagery. The need for accurate analysis of our environment is critical in areas such as disaster management, climate science, and agriculture. Given that satellite data is constantly changing due to both environment changes and human activities, it is essential for models to be resilient to distribution shifts over time. Additionally, due to the varying availability of data across different regions, it is crucial that these models generalize well across all regions, rather than being optimized only for data-rich areas.

Satellite imagery analysis presents unique challenges due to its variability and the high dimensionality of the data. The spatial and temporal resolution of satellite images can vary significantly, requiring models that can adapt to different scales and resolutions. In addition, satellite images may be taken at night, or with the presence of noise or atmospheric disturbances which pose further challenges for analysis. Traditional CNNs have made considerable strides in this domain, but their performance remain limited when dealing with OOD data.

ViTs offer a promising alternative due to their capabiltiy to capture long-range dependencies and model complex relationships within data. By leveraging self-attention mechanisms, ViTs can dynamically focus on different parts of an image, potentially leading to more robust feature representations. However, their application to satellite imagery, particularly in the context of OOD generalization, remains an area requiring further exploration.

The project conducts all experiments on the FMoW-Wilds dataset provided in the WILDS benchmark. The input to our algorithm is a 224 x 224 pixel RGB image, accompanied by metadata that includes the region and year of capture. We use a ViT model to output a predicted land use category from one of the 62 building or land use categories provided in the dataset.

2. Related Work

ViTs arose out of the successes of Transformers in NLP and applies a standard Transformer architecture directly to images. Several ViTs have revolutionized image recognition tasks by leveraging transformer architecture originally designed for natural language processing and applying them in either in conjunction with CNNs or in replacement of certain components of CNNs [1][4][7]. By splitting images into patches and treating these patches similarly to tokens in NLP applications, ViT models trained on mid-sized datasets such as ImageNet initially yielded modest accuracies slightly below those of comparable ResNets. However, when pre-trained on larger datasets, such as ImageNet-21k or JFT-300M, ViTs demonstrated superior performance, approaching or surpassing state-of-the-art results on multiple image recognition benchmarks, demonstrating they can match or exceed prior record performances by CNNs while also being less expensive to pre-train [4] [12].

Model regularization and augmentation techniques play a crucial role in impacting the performance of ViTs. There are two common data augmentation techniques: RandAugment, which randomly applies a set of image transformations such as rotation, translation, and color adjustments to increase the diversity of the training dataset, and Mixup, which blends pairs of images to create interpolated samples that enable the model to learn more robust representations. These data augmentation techniques improve ViT model performance across various image recognition benchmarks by providing richer training sets. Model regularization techniques such as Dropout and StochasticDepth also have significant impacts on ViT performance by discouraging overfitting [12]. We hope that applying these techniques will prove to increase performance in OOD tasks.

Domain adaptation and OOD detection are crucial for models that must deal with real-world data variability. Deep neural networks are able to learn powerful representations from large datasets, but they may not always generalize well when the training distribution differs from the test distribution [8]. The introduction of new datasets like FMoW, which contains over 1 million satellite images with temporal views, multispectral imagery, and metadata, help address OOD challenges by providing more diverse and comprehensive training data [2]. Minimizing performance degradation due to OOD also necessitates the use of domain adaptation algorithms to combat decline in model performance due to domain shift. DeepCORAL, one such algorithm, performs end-to-end adaptation in deep learning neural networks and exhibit state-of-the-art performance on on standard benchmarks [13]. Other research suggests that large-scale pre-trained transformers, in conjunction with few-shot outlier exposure setting, can significantly improve deep neural network performance on OOD tasks across different domains [5]. These findings highlight promising avenues for enhancing OOD detection, but there remains challenges in reliably detecting and handling OOD samples in real world scenarios and with ViTs.

Pre-training and fine-tuning strategies also help adapt transformers to specific tasks. BERT, a transformer-based model for NLP has performed well on tasks such as document classification, entity extraction, and question answering. Although BERT entirely excludes visual features, principles involved in BERT's pre-training and fine-tuning have influenced approaches to computer vision tasks [3]. Freezing certain layers during transformer fine-tuning helps deter overfitting and retain pre-trained knowledge, which has been effective in various transformer applications, including image recognition [9].

3. Methods

We evaluate two pre-trained checkpoints published alongside the original Vision Transformer paper [4]: the base (VIT-B/16) and large (VIT-L/16) sizes. Both models are trained on ImageNet [11] and ImageNet-21k [10] with Adam, a batch size of 4096 and L2 regularization with 0.1 decay. The base model has 12 layers and 12 attention heads and the large model has 24 layers and 16 heads. A baseline evaluation is also performed using DenseNet [6], a CNN with residual connections added by concatenating each layer's output to all downstream layers along the feature dimension.

Our ViT model handles images by breaking them into patches, adding position and class embeddings, and then passing them as tokens into the transformer encoders. These encoders are then connected to an MLP that builds a representation across all the patches. This representation is then fed into a class linear layer for classification. In our expirements, we will add the region and year metadata into the representation before passing it into the classifer in hopes that it will increase OOD performance. The ViT architecture can be seen in in 1.

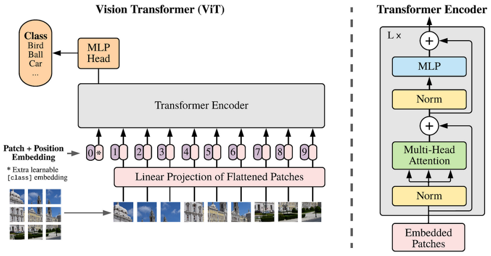

Figure 1. ViT architecture

One important augmentation we explore is Deep CORAL, a domain adaptation algorithm. Deep CORAL aligns the second order statistics of source and target feature distributions, which mitigates performance degradation that occurs when training and testing data distributions differ

[13].

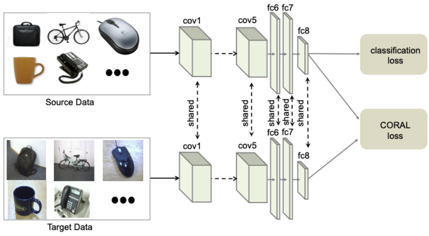

Figure 2. Sample Deep CORAL architecture on a CNN

Although originally developed for CNNs, the method can be adapted for ViTs. Specifically, after the ViT processes the image patches and constructs their representations through transformer encoders, we apply Deep CORAL loss to align the source and target domain feature representations [13]. This alignment reduces the gap between domains, allowing the ViT to maintain higher accuracy when applied to data from different distributions. Thus leveraging Deep CORAL enables our ViT to generalize across data distributions and perform better OOD.

Our project will build off the WILDS benchmark codebase. We will use the benchmark to fine-tune a pre-trained ViT model on our task. We will do this fine-tuning using an AdamW optimizer, and with a variety of of learning rates. Moreover, we will be able to add data augmentation to our data as to be more robust to OOD distributions. Because of our computation limitations, we will only work off results from 10 epochs. We trained on AWS instance with a Tesla T4 GPU.

Our added code will include the integration of the ViT model into the benchmark. Re-configuring the last linear layer to the appropriate number of classes. The option to freeze a specific number of layers. Modifying the dropout probabilities for the ViT. We will also add another algorithm by the name of 'MI' for metadata integration. This algorithm adds a linear layer to pass the region and year metadata through and concatenates this info to the image representation before passing it through the classifier. We hope that

We will also add tools that will allows for us to conduct qualitative analysis on our models. Such as printing confusion matrices and comparing the predictions of the models with the ground truth.

4. Dataset and Features

We will be using the FMoW-Wilds dataset that was provided in the WILDS benchmark. This dataset provides a comprehensive benchmark for evaluating models on land use classification and segmentation tasks. In this dataset, each input image is represented by a 224 x 224 pixel RGB image. Each input image is labeled with one of the 62 building or land use categories. Each example is additionally annotated with a metadata vector consisting of the year and the region (Africa, the Americas, Oceania, Asia, or Europe) it depicts.

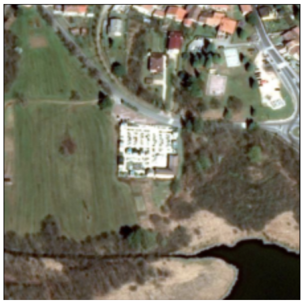

Figure 3. An example image from the train split. With metadata of Europe (1) and from 2012 (10).

During training, images will be augmented to create a more robust dataset and improve OOD performance. An example of such image can be seen in 4

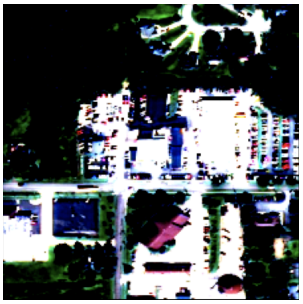

Figure 4. Augmented Image

Our dataset is broken into 5 splits. They are as follows:

1. **Training**: 76,863 images from the years 2002–2013.
2. **Validation (OOD)**: 19,915 images from the years from 2013–2016.
3. **Test (OOD)**: 22,108 images from the years from 2016–2018.
4. **Validation (ID)**: 11,483 images from the years from 2002–2013.
5. **Test (ID)**: 11,327 images from the years from 2002–2013.

These 5 splits represent how our models will be tests for OOD distribution. By training on a disjoint set from the OOD splits, we will effectively be able to test how time will affect our performance.

5. Experiments

We evaluate the out-of-distribution performance of our two ViT models and the CNN baseline on FMoW-WILDS. Each model is fine-tuned for 10 epochs with a seed of 0 on the FMoW training set, and we report test accuracies over the in-domain and out-of-domain splits on the worst-performing region, as in the WILDS guidelines. We also evaluate the effect of a number of simple ablations on ViT-B/16: dropout with $p = 0.1$ (VIT+DROPOUT-01), training only the last layer and freezing all others (VIT+LAST-FROZEN), and two algorithm-level changes: test-time transformation with DeepCoral (VIT+DEEPCORAL) and augmentation of the ViT image features with concatenation of the image metadata (VIT+METADATA). These primary results are summarized in 1, while loss graphs are presented for the training, ID validation and OOD validation spits in 5, 6, and 7 respectively.

6. Results

Model	Validation Acc.	Test Acc.	OOD Change	Training Time
DENSENET121	**0.578**	**0.463**	19.9%	**2.58hr**
VIT-B/16	0.569	0.436	23.4%	8.502hr
VIT+DROPOUT-0.1	0.561	0.430	23.4%	8.283hr
VIT+LAST-FROZEN	0.371	0.312	**15.9%**	4.134hr
VIT+DEEPCORAL	0.556	0.428	23.0%	8.283hr
VIT+METADATA	0.568	0.432	23.9%	8.181hr

Table 1. Worst-region accuracies on FMoW. OOD Change measures the percentage change in the reported test accuracy (out-of-domain) from the validation accuracy (in-domain).

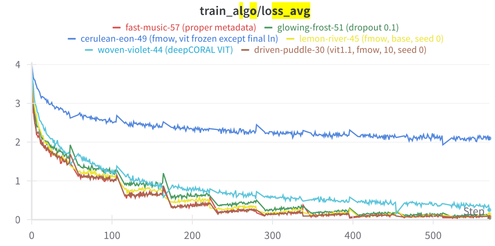

Figure 5. Training loss on FMoW. DENSENET121, VIT-B/16, VIT+DROPOUT-0.1, VIT+DEEPCORAL, VIT+METADATA are presented in yellow, brown, green, light blue and red respectively. VIT+LAST-FROZEN, presented in blue, fails to converge to the loss achieved by all other models.

6.1. Baseline Qualitative Analysis

We decided to look into the difference between our baseline and the VIT in more detail with an evaluation on the same OOD Test set of 22,108 images. We decided to look at the confusion matrices of the two models on their predicted class labels and analyze failure points of the two models.

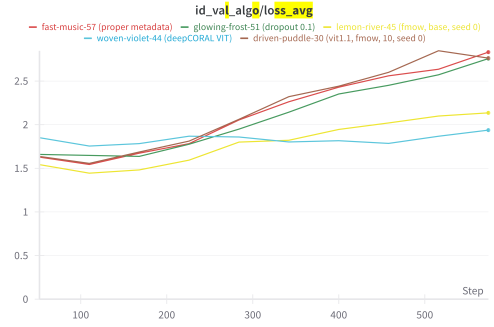

Figure 6. In-domain validation loss on FMoW.

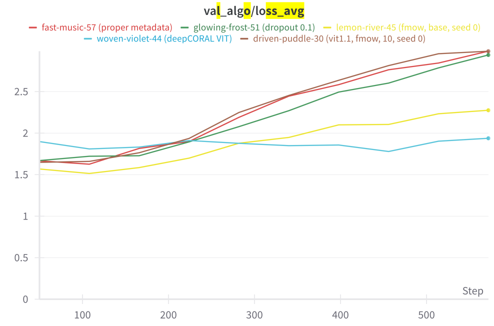

Figure 7. Out-of-domain validation loss on FMoW.

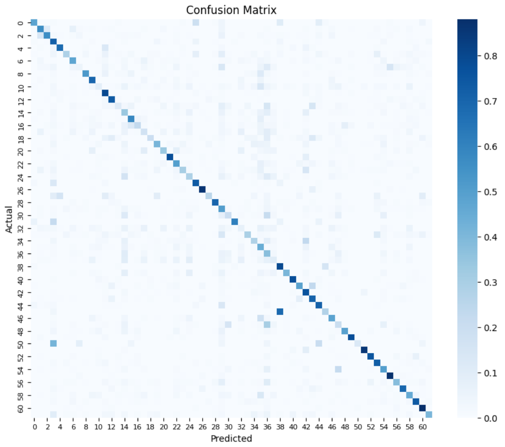

Figure 8. Baseline Confusion Matrix

The confusion matrix for our baseline can be seen in 8.

We can see from the confusion matrix that an image of a shipyard (45) was constantly confused as a multi-unit residential (30). As well, an image of a space facility (50) was regularly confused as a amusement park (3). There also appears to be some bias towards places of worship (36).

The top 5 and bottom 5 land-use classes by in-domain accuracy are presented in 2 and 3 respectively.

Class	Accuracy (%)	Count
Wind Farm (0)	87.97	399
Toll Booth (55)	87.22	352
Interchange (26)	85.91	220
Stadium (51)	85.22	379
Crop Field (11)	79.60	1137

Table 2. Top 5 classes by baseline accuracy.

Class	Accuracy (%)	Count
Border Checkpoint (7)	4.95	101
Office Building (32)	6.89	421
Construction Site (10)	8.41	214
Police Station (37)	10.43	345
Debris or Rubble (13)	10.84	83

Table 3. Bottom 5 classes by baseline accuracy.

Class	Accuracy (%)	Count
Interchange (26)	82.27	220
Toll Booth (55)	87.22	352
Crop Field (11)	85.91	1137
Solar Farm (49)	85.22	311
Stadium (51)	79.60	379

Table 4. Top 5 classes by ViT accuracy.

Class	Accuracy (%)	Count
Construction Site (10)	7.84	214
Office Building (32)	7.60	421
Debris or Rubble (13)	9.64	83
Police Station (37)	14.20	345
Fire Station (17)	14.74	312

Table 5. Bottom 5 classes by ViT accuracy.

6.2. ViT Qualitative Results

The confusion matrix for our baseline can be seen in 9.

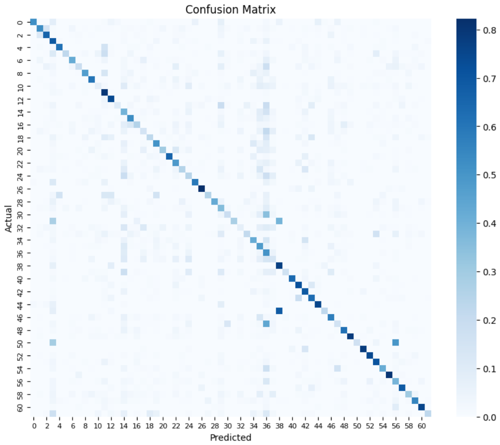

Figure 9. VIT Confusion Matrix

We can see from the confusion matrix, that there is still a some confusion of wrongly classifying a shipyard (45) as a multi-unit residential.

We can see the the top 5 classes based on accuracy for our VIT model in 4.

Furthermore, we can see the bottom 5 classes based on accuracy for the VIT model in 5.

7. Discussion

7.1. Quantitative Analysis Comparison

Overall, we find that standard vision transformers fail to outperform CNNs on the FMoW task, despite their much larger computational overhead, during both training and inference. The DenseNet baseline achieves the best ID validation accuracy, OOD test accuracy and training time, as well as performing best on our direct metric of domain adaptation ability—the performance drop between ID and OOD splits. Furthermore, most of the ViT variant ablations showed comparatively minimal effect relative to the base VIT-B/16, except for VIT+LAST-FROZEN, which reports the lowest final accuracies by a wide margin. This is likely because when all low-level layers are frozen, the residual number of trainable parameters does not provide enough capacity for the ViT to properly learn the task during fine-tuning.

Since all models achieve training convergence and similar final training losses, however, we cannot attribute the drop in ViT performance otherwise to a lack of capacity to learn the task; conversely, there is significant evidence of overfitting on most of the ViT models indicated by a marked rise in validation loss for both ID and OOD settings. It is most likely that the benefits of self-attention observed in prior work with ViTs failed to be realized on this dataset, though further study would be needed to claim more precise causes.

One clear positive result is that DeepCoral improves generalization for ViTs, both in and out of domain. The validation loss trajectories for VIT+DEEPCORAL, marked in light blue in figures 6 and 7, limit overfitting to a greater extent than even the DenseNet baseline. In view of the algorithm's relative simplicity, it is encouraging to see its effectiveness confirmed on ViT's.

7.2. Qualitative Analysis Comparison

Both of our models showed major confusion between images of shipyards (45) and multi-unit residential (30). Yet, the ViT model doesn't seem to confuse space facilities for amusement parks as much. As well, both models continue to show bias towards places of worship (36), while the ViT demonstrating a stronger bias towards this class. We can attribute this decrease in bias towards places of worship in the baseline, as a possible reason why the baseline performs better for OOD.

The top and bottom classes by accuracy for both models indicate that both models achieve high accuracy on classes like toll booths (55) and crop fields (11), and both perform poorly on classes such as construction sites (1) and office buildings (32). One important difference is that baseline accuracy is significantly lower for its worst performing classes compared to the ViT. Yet, the baseline on average performs better. This indicates that the baseline model might be disproportionately optimizing for some classes to achieve better aggregate performance. Some further analysis could be conducted to see what learned features are differ between the models for Border Checkpoint (7) examples. Since it is the lowest for the baseline, but was not part of the lowest for the ViT model.

8. Conclusion & Future Work

In this work we analyze the performance of vision transformers (ViTs) on a satellite imagery classification task in both in-domain and out-of-domain settings. We determine that the standard vision transformer fails to achieve significant performance improvement over a pretrained DenseNet. These findings counter the emergent narrative in the literature that ViTs are more robust learners than their CNN counterparts, and again highlight the need for more granular studies on the robustness/transferability of ViT learning.

The comparative analysis revealed that while ViTs offer more advanced representational capabilities, they can still struggle with domain adaptation challenges similar to CNNs. Integrating Deep CORAL and metadata provided incremental improvements, but could not fully address performance degradation in OOD data. This suggests that the robustness of ViTs is not as universal as previously thought and is highly dependent on specific architectural adaptations and the nature of the task.

Future work should focus on addressing the identified weak points in ViTs. In [14], for example, static position embeddings are found to significantly inhibit domain adaptation performance compared to learned variants, while a lack of residual feed-forward connections in the standard architecture is also identified as a week point. In follow-up study, we would analyze whether new adaptations akin to their proposed RVT architecture can be developed to address these weak points specifically in our domain of satellite imagery.

Ultimately while ViTs demonstrate potential for high performance image classification, their robustness and transferability continue to require further investigation.

9. Contributions & Acknowledgements

The original codebase for WILDS can be found at https://github.com/p-lambda/wilds. Any changes can be found on https://github.com/Adriatogi/wilds

Adrian Gamarra Lafuente: AWS instantiation. Code for ViT integration and its different possible configurations. Metadata integration. Training and evaluation of different models. Code for Qualitative analysis and results. Weights and biases logging. Parts of the final report.
Mai Hoang: Parts of final report. Related works/literature review.
Nathan Kim: Initial specification of project scope (domain adaptation), base models and datasets. Tabulation + discussion of final quantitative results in report.

References

[1] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan. Vision transformers for remote sensing image classification. 2021. 2

[2] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee. Functional map of the world, 2018. 2

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 2

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. 1, 2

[5] S. Fort, J. Ren, and B. Lakshminarayanan. Exploring the limits of out-of-distribution detection, 2021. 1, 2

[6] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks, 2018. 2

[7] M. Kaselimi and N. D. A. D. Athanasios Voulodimos, Ioannis Daskalopoulos. A vision transformer model for convolution-free multilabel classification of satellite imagery in deforestation monitoring. 2023. 2

[8] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2021. 1, 2

[9] J. Lee, R. Tang, and J. Lin. What would elsa do? freezing layers during transformer fine-tuning, 2019. 2

[10] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 2

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2

[12] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer. How to train your vit? data, augmentation, and regularization in vision transformers, 2022. 2

[13] B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation, 2016. 2, 3

[14] J. Yang, J. Liu, N. Xu, and J. Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation, 2021. 6