CS330 Final Report: Geography-Aware Few-Shot Remote Sensing Scene Classification

Ethan Hellman
Stanford University, Computer Science
hellman1@stanford.edu

Fall 2023

Abstract

This paper delves into the novel integration of multi-modal models, particularly RemoteCLIP, with Model-Agnostic Meta-Learning (MAML) in the context of the Functional Map of the World (fMoW) dataset, to address the challenge of few-shot learning in remote sensing scene classification. Central to this study is the exploration of how multi-modal models, which synergize visual and semantic data, can be effectively adapted to sparse data scenarios prevalent in remote sensing.

The RemoteCLIP model, selected for its advanced vision-language processing capabilities, is at the forefront of this study. It is based on the CLIP framework and employs a ResNet-50 image encoder. The model's intrinsic ability to process language and visual information enables it to excel in various tasks, including zero-shot image classification and object counting. To adapt and evaluate RemoteCLIP, the fMoW dataset was employed, comprising over 1 million images from more than 200 countries, each enriched with detailed bounding box annotations across 63 diverse categories. This dataset's extensive global coverage and rich metadata make it an ideal resource for training and testing the model.

Integrating MAML with RemoteCLIP was a strategic decision aimed at enhancing the model's few-shot learning capability. MAML's flexibility and effectiveness in rapid adaptation to new tasks complement RemoteCLIP's multi-modal framework, creating a robust system for remote sensing scene classification.

The study's experimental phase involved zero-shot geography probing, semantic injection, and few-shot fine-tuning, each designed to test and enhance RemoteCLIP's geographic knowledge and classification accuracy. The results from these experiments are noteworthy, demonstrating that incorporating semantic metadata with visual data significantly improves the model's performance in few-shot learning scenarios. Particularly, the model's ability to understand and integrate geographic information into its learning process was substantially enhanced. This finding is the key contribution of our study, demonstrating that semantic metadata, when skillfully integrated with visual data, can markedly enhance the performance of few-shot learning models. While this study is a preliminary investigation of these behaviors, the results shown should merit further exploration of these techniques such that we can better leverage remote sensing data to create intelligent machinery for critical applications in data-sparse regimes.

1 Introduction

In recent years, the field of machine learning has experienced significant advancements, with vision emerging as a notably dynamic area. The evolution from convolutional networks to sophisticated deep learning frameworks has propelled image models to reach, and in some instances surpass, human-level performance across diverse tasks [6]. The advent of ImageNet was a landmark moment, highlighting the need for large-scale data to train Artificially Intelligent systems, thus heralding a new era in machine learning [4].

Today's most advanced models are trained on datasets far larger than those envisioned by the field's early pioneers. However, a critical challenge persists: learning from a limited supply of data, known as "few-shot" learning. While humans excel at this intuitively from a young age, it remains a complex problem for machines [7]. Domains like language modeling may face fewer hurdles in few-shot learning, but in fields like remote sensing, the scarcity of high-quality, labeled data is a significant challenge [20].

Remote sensing data, often expensive and time-consuming to collect and label, offers invaluable insights for critical applications such as climate change monitoring and assessing environmental impacts [16]. This underscores the urgency and importance of developing efficient learning methods for sparse data.

This paper explores new directions in learning from limited data within the remote sensing context. We note that multi-modal models, which integrate visual and semantic data, are increasingly adept at image classification [12]. Additionally, satellite imagery is often accompanied by rich metadata, presenting an opportunity to leverage this information. This study investigates whether utilizing semantic metadata to fine-tune advanced remote sensing multi-modal models can enhance their few-shot learning capabilities. Our aim is to assess if this approach allows models to effectively learn from sparse data by integrating both visual and semantic cues.

Positioned at the intersection of few-shot learning, machine learning, and multi-modal models, this study contributes to the advancement of efficient learning methodologies in remote sensing. The ability to maximize the value of limited data in this field can have significant and far-reaching impacts, particularly in pressing global issues such as environmental monitoring and climate change.

2 Related Works

2.1 Remote Sensing Scene Classification

Remote sensing scene classification has significantly evolved with the integration of deep learning techniques. The field has seen innovative approaches in feature selection, attention mechanisms, and network architecture adaptation. For instance, Zou et al. in "Deep Learning Based Feature Selection for Remote Sensing Scene Classification" [24] demonstrated the effectiveness of deep-learning-based feature selection in high-resolution satellite image classification.

Another noteworthy development is the introduction of the Multi-Branch Local Attention Network by Chen et al. in "Remote Sensing Scene Classification via Multi-Branch Local Attention Network" [2], emphasizing the role of attention mechanisms in complex scene interpretation. Additionally, the exploration of CNN architectures, as highlighted by Broni-Bediako et al. in "Searching for CNN Architectures for Remote Sensing Scene Classification" [1], and the innovative use of relation-attention models in Wang et al.'s "Relation-Attention Networks for Remote Sensing Scene Classification" [19], signify the ongoing advancements in this domain.

2.2 Few-Shot Learning & Meta Learning

Few-shot learning is revolutionizing the way models generalize from limited data. A landmark study by Radford et al., "Learning Transferable Visual Models From Natural Language Supervision" [12], introduced the CLIP model, blending language and vision models to understand and classify images with minimal training data. This approach is especially pertinent in remote sensing, where labeled data can be scarce. Other significant contributions include Vinyals et al.'s "Matching Networks for One Shot Learning" [17], which introduced a novel approach for one-shot learning, and Snell et al.'s "Prototypical Networks for Few-shot Learning" [15], which emphasized the role of metric learning in few-shot classification tasks.

2.3 Few-Shot Remote Sensing Scene Classification

Few-shot remote sensing scene classification, an emerging field crucial for interpreting limited data, is being transformed by advanced meta-learning techniques and innovative approaches. Zhang et al. in "RS-SSKD: Self-Supervision Equipped with Knowledge Distillation for Few-Shot Remote Sensing Scene Classification" [23] present a novel two-branch network that effectively utilizes self-supervision and knowledge distillation, enhancing classification performance in data-scarce scenarios. This approach is further complemented by Xing et al.'s "Learning to Cooperate: Decision Fusion Method for Few-Shot Remote-Sensing Scene Classification" [22], which introduces a decision fusion model using pre-trained feature extractors for enhanced feature discrimination.

Li et al. in "SCL-MLNet: Boosting Few-Shot Remote Sensing Scene Classification via Self-Supervised Contrastive Learning" [10] integrate self-supervised contrastive learning with few-shot classification algorithms, facilitating effective learning from a limited number of annotated samples. A significant contribution in this area is also seen in "Meta-Learning for Few-Shot Land Cover Classification" [13] by Rußwurm et al., which demonstrates the application of meta-learning to land cover classification, showcasing the model's adaptability to new tasks with minimal examples.

Additionally, Li et al.'s "RS-MetaNet: Deep Metametric Learning for Few-Shot Remote Sensing Scene Classification" [9] introduces a metametric learning

approach that shifts the focus from sample-level to task-level learning, significantly enhancing the model's generalization capabilities. These studies collectively represent a significant stride in few-shot remote sensing scene classification, showcasing the potential of combining meta-learning, self-supervision, and advanced neural network architectures to address the challenges posed by limited training data in remote sensing.

2.4 Multi-modal Models in Remote Sensing

In the evolving field of remote sensing, the integration of multi-modal models, particularly CLIP and Vision Transformers (ViT) [5], is revolutionizing the analysis of satellite and aerial imagery. The study "Learning Transferable Visual Models From Natural Language Supervision" [12] by Radford et al. highlights CLIP's effectiveness in bridging visual content with textual descriptions, a pivotal advancement for classifying and interpreting remote sensing imagery. Vision Transformers, known for treating images as sequences of patches, offer a novel approach for global feature understanding, as seen in various studies, enhancing scene analysis in remote sensing.

The potential of multi-modal approaches in remote sensing is further illustrated in works like "Zero-Shot Multi-Modal Artist-Controlled Retrieval and Exploration of 3D Object Sets" [14] by Schlachter et al. This study demonstrates the versatility of multi-modal models in understanding complex datasets. Additionally, the survey "Large-scale Multi-modal Pre-trained Models: A Comprehensive Survey" [18] by Wang et al. underscores the growing importance of these models in various domains, including remote sensing. Furthermore, Lee et al.'s creation of the "DialogCC: Large-Scale Multi-Modal Dialogue Dataset" [8] opens possibilities for training remote sensing models on varied data, including dialogues and imagery, potentially enhancing scene interpretation.

Overall, the adoption of multi-modal models like CLIP and ViT is leading to more sophisticated and accurate remote sensing applications, significantly enhancing our ability to analyze and understand Earth's landscapes through advanced satellite and aerial imagery.

2.5 Remote Sensing Foundation Models

These observations are further seen in the development of foundation models specific to remote sensing, a potential game-changer for the field. Notable works include "DINO-MC: Self-supervised Contrastive Learning for Remote Sensing Imagery with Multi-sized Local Crops" [21] by Caron et al., which explores self-supervised learning for remote sensing imagery, and "SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery" [3] by He et al., focusing on pre-training transformers for diverse satellite imagery types. These studies highlight the potential of tailored foundation models in enhancing the analysis and interpretation of remote sensing data.

2.6 Summary

The convergence of few-shot learning, multi-modal learning, and domain-specific modeling is significantly reshaping remote sensing. This synergy enhances data interpretation by integrating the comprehensive analysis of multi-modal learning with the adaptability of few-shot learning. Domain-specific models further refine this approach, offering tailored analyses for remote sensing data's unique demands.

These advancements represent a transformative shift, promising more intelligent, efficient, and contextually aware remote sensing technologies. This integration heralds a new era in the field, leveraging the rich potential of remote sensing data to deepen our understanding of the Earth's landscapes and phenomena.

3 Methods

3.1 RemoteCLIP

In this study, we implement the RemoteCLIP model as introduced in "RemoteCLIP: A Vision Language Foundation Model for Remote Sensing." RemoteCLIP is a pioneering model that integrates vision and language, designed to address the limitations of traditional remote sensing models focused primarily on low-level features.

3.1.1 Model Selection and Architecture

RemoteCLIP was selected due to its ease of implementation, robust training on an expansive dataset, and promising results in preliminary evaluations. The model is structured on the CLIP framework, utilizing a ResNet-50 image encoder with 38M parameters, chosen for its computational practicality and efficiency.

3.1.2 Key Features of RemoteCLIP

1. Capability in Retrieval and Zero-Shot Applications: RemoteCLIP excels in tasks such as zero-shot image classification, few-shot classification, image-text retrieval, and object counting. Its ability to understand language enables it to perform these tasks effectively, making it suitable for diverse downstream applications in remote sensing.

2. Training Methodology: The model employs the CLIP strategy, known for its excellent generalization ability in vision-language tasks. It optimizes the InfoNCE loss function to align image-text pairs and distinguish mismatches. This process involves a large-scale dataset to encode images and texts into latent representations.

3. Data Scaling via Annotation Unification: A significant innovation in RemoteCLIP is its approach to scaling data using Annotation Unification.

This method expands the training dataset by converting object bounding box annotations into natural language captions, overcoming data scale limitations.

4. Optimization and Training Data: The final training data comprises 165,745 images, each with five captions, resulting in 828,725 image-text pairs. The optimization process is based on the ITRA codebase, developed from OpenCLIP, and includes automatic mixed-precision and the Adam optimizer.

3.1.3 Application in Remote Sensing

Given its recent introduction and the uncharted territory in its application, RemoteCLIP's implementation in this study is exploratory. The model's unique ability to understand language and process visual features with rich semantics offers an advanced approach to remote sensing scene classification, addressing the challenges posed by limited annotated data and enhancing accuracy in various tasks.

In summary, the choice of the RemoteCLIP model, particularly its ResNet-50 variant, aligns with our objective to explore advanced, domain-specific solutions in remote sensing, leveraging its multi-modal learning capabilities for efficient and effective scene classification.

3.2 MAML

In this study, we have chosen to incorporate the Model-Agnostic Meta-Learning (MAML) algorithm, as presented in "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks" by Finn et al. (2017). MAML was selected for its ease of implementation and its proven track record in the field, especially for its effectiveness in few-shot learning scenarios.

3.2.1 Key Considerations for Choice of MAML:

1. Proven Effectiveness: MAML is widely recognized for its strong performance in meta-learning, particularly in rapidly adapting to new tasks with limited data. This aspect is pivotal for remote sensing applications, where diverse and sometimes scarce data sets are common.

2. Simplicity and Flexibility: The straightforward implementation process of MAML made it a practical choice for our project. Its flexibility allowed for seamless integration with the RemoteCLIP model, which was crucial for enhancing the model's adaptability to a range of remote sensing scenes.

3. Model-Agnostic Nature: As a model-agnostic algorithm, MAML can be applied to various architectures, including the complex structure of RemoteCLIP. This quality ensured that we could enhance RemoteCLIP's learning capability without extensive modifications to its architecture.

3.2.2 Application in Remote Sensing

MAML's ability to learn quickly from a small number of examples makes it particularly suitable for our remote sensing classification task. In a field where annotated data may be limited, MAML's approach to learning provides a significant advantage. It enables the RemoteCLIP model to adapt rapidly to new tasks and scenarios, a key requirement for effective remote sensing analysis.

In summary, the integration of MAML into our study was a strategic decision to leverage its meta-learning capabilities and enhance the performance of the RemoteCLIP model in classifying remote sensing scenes. This combination aims to develop a system that is not only robust and adaptable but also capable of effectively handling the diverse challenges in remote sensing scene classification.

3.3 Functional Map of the World Dataset

In our study, we utilized the "Functional Map of the World" (fMoW) dataset, specifically chosen for its comprehensive collection of satellite images tailored to advance machine learning models in remote sensing. This dataset is instrumental in classifying the functional purpose of buildings and land use from satellite imagery.

3.3.1 Dataset Composition

The fMoW dataset includes over 1 million images from more than 200 countries. Each image features at least one bounding box annotation, classified into one of 63 categories, including a "false detection" category for content not fitting the other 62 classifications. This extensive and diverse categorization is essential for our study's focus on a wide range of remote sensing scenarios.

3.3.2 Unique Features

Characterized by its global diversity, the fMoW dataset is enriched with detailed metadata accompanying each image. This metadata includes vital information such as location, sun angles, and physical sizes of the imaged objects or areas, providing essential context for accurate predictions and classifications.

3.3.3 Dataset Version Utilized

For our experiment, the fMoW-rgb version was selected, comprising JPEG-compressed images. This version aligns with our computational resources and provides the necessary visual data for analysis. The fMoW-rgb includes RGB channels extracted and saved in JPEG format, offering a practical yet comprehensive dataset for our remote sensing study.

3.3.4 Rationale for Dataset Selection

The selection of the fMoW dataset, particularly the fMoW-rgb version, was driven by its expansive size and diversity, making it an invaluable resource for robust machine learning model development in remote sensing. Crucially, the RemoteCLIP model had not been trained on this dataset, ensuring that our model adaptation was challenged with entirely unseen data. This aspect is vital for evaluating the model's adaptability and generalization capability in new and diverse remote sensing contexts.

In summary, the "Functional Map of the World" dataset's extensive coverage and rich metadata make it an ideal choice for this study, providing a varied and realistic platform for training and evaluating the RemoteCLIP model's efficacy in remote sensing scene classification.

3.4 Task Selection

To fine-tune RemoteCLIP for few-shot learning, we transformed the FMoW dataset into a few-shot dataset, creating tasks that involve scene classification across different countries. The objective is for the model to quickly adapt to a new country based on a small set of labeled images and achieve high-accuracy scene classification after a few fine-tuning iterations. Accordingly, we partitioned the FMoW dataset by country and allocated them to `train`, `val`, and `test` subsets.

A significant challenge was the dataset's imbalance: some countries have tens of thousands of labeled images, whereas others have as few as one. This discrepancy, along with the socio-economic narratives it may imply, was not addressed by the original authors. To establish equitable task distribution, we implemented a task allocation algorithm that excluded countries with fewer than 32 images and capped the maximum at 20,000 images per country. This process excluded 29 countries and reduced the dataset by 75,402 images, leaving 324,053 images for training. Additionally, ensuring a balanced distribution of tasks and images and uniform representation of categories across data splits was paramount. We developed a novel allocation algorithm to meet these constraints and achieve an optimal balance of tasks, images, and categories. The essence of this method is presented in **Algorithm 1**, and a detailed abstraction of the full implementation is provided in **Appendix A**, alongside descriptions of supporting methods.

Algorithm 1 Set Selection Based on Weighting

function SELECTSETBASEDONWEIGHTS(sets, set_targets, category_sets, country_category)
    train_weight $\leftarrow \left(\frac{\text{SIZE}(sets['train'])}{set\_targets['train']}\right) + \left(\frac{\text{SIZE}(category\_sets['train'])}{set\_targets['train\_categories']}\right)$
    val_weight $\leftarrow \left(\frac{\text{SIZE}(sets['val'])}{set\_targets['val']}\right) + \left(\frac{\text{SIZE}(category\_sets['val'])}{set\_targets['val\_categories']}\right)$
    test_weight $\leftarrow \left(\frac{\text{SIZE}(sets['test'])}{set\_targets['test']}\right) + \left(\frac{\text{SIZE}(category\_sets['test'])}{set\_targets['test\_categories']}\right)$
    missing_categories $\leftarrow$ COMPUTEMISSINGCATEGORIES(*category_sets*)
    if country_category in missing_categories['train'] **then**
        train_weight $\leftarrow$ ADJUSTWEIGHTFORMISSINGCATEGORY(*train_weight*)
    end if
    if country_category in missing_categories['val'] **then**
        val_weight $\leftarrow$ ADJUSTWEIGHTFORMISSINGCATEGORY(*val_weight*)
    end if
    if country_category in missing_categories['test'] **then**
        test_weight $\leftarrow$ ADJUSTWEIGHTFORMISSINGCATEGORY(*test_weight*)
    end if
    return SELECTMINWEIGHTSET(train_weight, val_weight, test_weight)
end function

Data Split	**Value**
Total Images	324053
Total Countries	139
Train Images	213313
Train Images Percent	0.658
Train Countries	95
Train Countries Percent	0.683
Val Images	54514
Val Images Percent	0.168
Val Countries	22
Val Countries Percent	0.158
Test Images	56226
Test Images Percent	0.174
Test Countries	22
Test Countries Percent	0.158

Table 1: Data Splits for FMoW Dataset

As can be see in **Table 1**, our algorithm was able to surprisingly allocate countries and images effectively to achieve a rough 70/15/15 split of the data.

4 Experiments

4.1 RemoteCLIP Zero-Shot Geography Probing

Our preliminary experiments aimed to evaluate the geographical knowledge encoded within the RemoteCLIP model, specifically its capacity to discern geographic information at varying granularities, including hemisphere, continent, and country levels, integrated within the interplay between the text and image encoders.

To this end, we employed a zero-shot information retrieval approach where the model was tasked with matching images to their geographic metadata without prior explicit training on these specific tasks. Queries were derived from the metadata alone or in combination with the hand-crafted prompts utilized during RemoteCLIP's training phase (e.g., "Europe", "Iran", or more descriptive forms like "a satellite image of the North Eastern Hemisphere", "a satellite image of Africa", "a satellite image of Mexico"). These queries were processed by the text encoder, and each corresponding image was encoded through the image encoder.

The underlying hypothesis was straightforward: if the text and image encoders have successfully internalized geographic information, then the nearest neighbor in the embedding space for any given image should be the text description accurately reflecting its geographic context. To benchmark this geographic retrieval performance, we also assessed the model's capability in class retrieval tasks across different object categories.

4.2 RemoteCLIP Geographic Semantic Injection

The objective of our subsequent experiments was to determine whether the RemoteCLIP network could internalize geographic knowledge. This was approached by fine-tuning the network with semantically enriched metadata to align the representations between the image and text encoders. We adopted a hybrid training goal that merged contrastive estimation with classification. Specifically, geographic metadata was utilized to generate descriptive prompts for each satellite image—e.g., "a satellite image of Australia in the South Eastern hemisphere of Oceania." These prompts, alongside their corresponding images, were processed by their respective encoders, and the resulting embeddings were concatenated. This combined feature vector was then input into a two-layer neural network to obtain logits, which were used to categorize the images using a softmax classifier. The total loss comprised the sum of the embedding distance and the classification error. For this experiment, we fine-tuned the final layers of both encoders while keeping the preceding layers fixed.

4.2.1 Normalization via Information Dropout

To ensure the RemoteCLIP model could effectively discern geographic details at various scales, we introduced a normalization technique akin to information

dropout. For any piece of geographic data, we programmed a probability p that it would be omitted from the textual prompt. Thus, prompts would alternate between being fully specified, such as "A satellite image of Australia in the South Eastern hemisphere of Oceania," and partially redacted, like "A satellite image of ___ in the South Eastern hemisphere of ___." This strategy was designed to compel the model to consider all levels of geographic detail and avoid over-reliance on specific features.

4.3 RemoteCLIP Few-Shot Fine-tuning for Geography-Aware Remote Sensing Scene Classification

Building on our semantic injection experiments, we aimed to adapt RemoteCLIP to a few-shot learning scenario, aligning with the central theme of this study. To achieve this, we employed the Model-Agnostic Meta-Learning (MAML) framework, transforming the FMoW dataset into a few-shot compatible format. Our fine-tuning process involved several nuanced scenarios.

Integration with the PyTorch MAML framework, as outlined in [11], necessitated the creation of two classes: `CLIPModel` to work with RemoteCLIP and `LinearClassifier` for the linear classifier applied to embeddings from RemoteCLIP's image and text encoders. We limited fine-tuning to the last visual layer of ResNet-50 and the residual block of the transformer text encoder. Due to computational constraints, we subsampled 500 out of approximately 2500 tasks for training and 250 out of roughly 750 tasks for validation in each epoch. Our experimental setup was consistently `5-way`, `5-shot`, evaluated on 5 query images per class. The meta learning rate was set at 0.001, while the inner loop's update learning rate was 0.01. In scenarios with distinct learning rates for CLIPModel and LinearClassifier, the former's learning rate was 100 times smaller. Each batch of tasks in the inner loop underwent 10 update steps.

The conducted experiments, designed to probe various aspects of few-shot fine-tuning, included:

- RemoteCLIP fine-tuned with Geoprompting + Dropout ($p = 0.2$), shared learning rate
- RemoteCLIP fine-tuned with Geoprompting + No Dropout ($p = 0.0$), shared learning rate
- RemoteCLIP fine-tuned without Geoprompting ("a satellite image"), shared learning rate
- RemoteCLIP fine-tuned with Geoprompting + Dropout ($p = 0.2$), separate learning rates
- RemoteCLIP fine-tuned with Geoprompting + No Dropout ($p = 0.0$), separate learning rates
- RemoteCLIP fine-tuned without Geoprompting ("a satellite image"), separate learning rates

- RemoteCLIP fine-tuned without prompting (image-only), separate learning rates

These experiments were meticulously designed to evaluate the impact of our information dropout normalization on few-shot performance, the differential effect of tuning learning rates between CLIPModel and LinearClassifier, and the overall influence of textual prompting in the fine-tuning process.

5 Results & Discussion

5.1 RemoteCLIP Zero-Shot Geography Probing

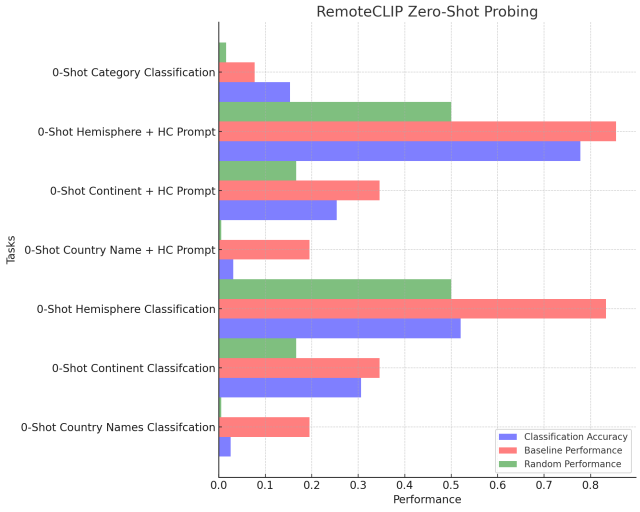


Figure 1: RemoteCLIP Geography Knowledge Analysis

Figure 1 presents the performance of the RemoteCLIP model across various geographical probing tasks. Notably, the model does not exhibit a clear trend of geographic awareness. It underperforms relative to the baseline in all tasks except categorical classification. The latter serves as a control to verify RemoteCLIP's compatibility with satellite imagery; indeed, the model demonstrates an ability to exceed baseline performance in scene classification, despite not being trained on the FMoW dataset.

This outcome suggests that while RemoteCLIP can generalize to satellite data to some extent, its geographical knowledge, particularly at granular levels, is limited. Additionally, the results highlight a significant skew in the data, which underscores the importance of a nuanced approach when assessing the model's performance. The insights gleaned here have shaped the subsequent experimental design, ensuring a more refined evaluation of the model's capabilities in geographically-informed tasks.

5.2 RemoteCLIP Geographic Semantic Injection

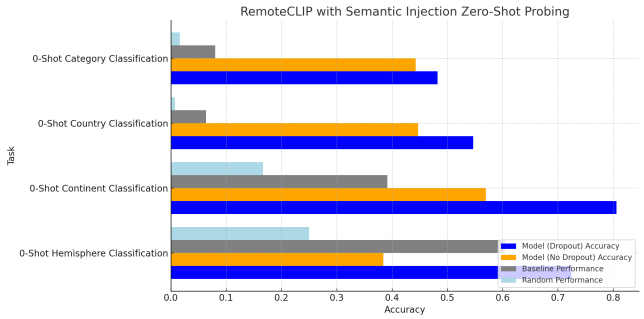


Figure 2: RemoteCLIP with Semantic Injection Geography Knowledge Analysis

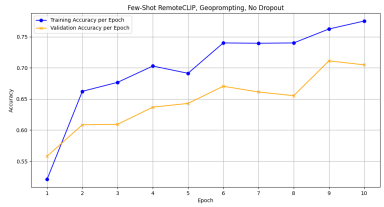
Figure 2 illustrates the impact of semantic injection on RemoteCLIP's performance in zero-shot geographical tasks. The model's marked improvement in this setting is particularly noteworthy given that, during testing, only images were provided—no accompanying geographic data. This suggests that the model has effectively learned to recognize geographic features or features correlated with geographic semantics.

The data indicate that for the majority of tasks, the model—both with and without dropout—surpasses the baseline performance, with some tasks showing substantial gains. This finding confirms our hypothesis that the model can be trained to develop a sense of geographic awareness within our set constraints. A crucial observation is the consistent outperformance of the dropout-regularized model over its counterpart, implying the effectiveness of our dropout strategy in facilitating the model's ability to discern geographic information. Additionally, the enhanced performance on classification tasks post-training attests to the benefits of the additional semantic context provided during training.

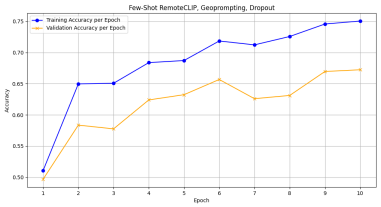
5.3 RemoteCLIP Few-Shot Finetuning for Geography-Aware Remote Sensing Scene Classification

In the unified learning rate regime for both the `CLIPModel` and `LinearClassifier`, we encountered training instability leading to a collapse in accuracies, rendering the train and test results non-reportable.

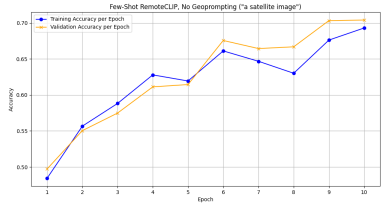
Conversely, the results from other training scenarios are revealing and promising.



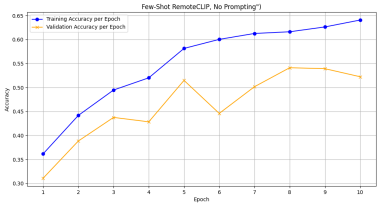
(a) Geoprompting + No Dropout



(b) Geoprompting + Dropout



(c) No Geoprompting



(d) No Prompting

Figure 3: Training Curves for all functioning models over 10 Epochs

Model	**Accuracy**
Geoprompting, No Dropout	0.2459
Geoprompting, Dropout	0.2817
No Geoprompting	0.2462
No Prompting	0.1173

Table 2: Accuracy of Few-Shot Finetuned Models

As demonstrated in **Figure 3**, the remaining models exhibit a positive trajectory in adapting to few-shot settings, with **Table 1** providing a quantified overview of their performance. Notably, even within the abbreviated span of 10 epochs and with a subsampled dataset, the models attain substantial accuracy. This is particularly promising, considering that models in traditional studies are typically trained for longer periods and with full datasets. The absence of a plateau in the training and validation curves suggests that further improvements might be achievable with more extensive training. This is highly encouraging for further studies that aim to push adaptation further

Table 1 reveals that the models fine-tuned with geoprompting—both with and without dropout—achieve accuracies significantly above the 'No Prompting' baseline, confirming the beneficial impact of prompts on few-shot learning performance. The 'Geoprompting, Dropout' model outperforms its counterparts, signifying the effectiveness of the dropout strategy in enhancing few-shot learning. Most notably, all models fine-tuned with prompts significantly surpass the 'No Prompting' model that relies solely on visual input. These results

corroborate the premise that incorporating metadata can substantially bolster few-shot learning capabilities in multi-modal remote sensing classifiers, meriting further in-depth investigation.

5.4 Limitations and Future Work

This investigation has laid the groundwork for significant strides in multi-modal few-shot learning, yet it recognizes the boundaries set by resource constraints. The compute-intensive nature of applying a meta-learning framework to a large-scale model like RemoteCLIP—with its extensive parameter set and the consequent second-order gradients required by MAML—imposed substantial demands on cloud computing resources. These limitations inherently restricted the scope of hyperparameter exploration and the feasibility of employing larger image encoders, such as those in the `ViT` series.

Our experience revealed the necessity for distinct learning rates between the `LinearClassifier` and `CLIPModel`, and we observed training instabilities associated with the text encoder over extended periods. This suggests that a static learning rate may not suffice for long-term training stability, indicating a fertile area for future research to delve into adaptive learning rate strategies and differential rates for image and text encoders.

Furthermore, the solitary nature of this project's execution hints at the vast potential that could be unlocked with collaborative efforts. The encouraging initial results merit the attention of the broader research community, opening avenues for collective exploration into the variables affecting model performance. Future endeavors should aim to expand on the number of shots, query samples, and epochs, alongside a more granular adjustment of learning rates. With collaborative synergy, there's a promising horizon for advancing the capabilities of multi-modal few-shot classifiers and enhancing their practical utility in remote sensing applications.

6 Conclusion

This study represents a advancement in the domain of remote sensing, demonstrating the untapped potential of multi-modal models within few-shot learning frameworks. By integrating RemoteCLIP with the MAML approach and applying it to the fMoW dataset, we have illustrated that semantic metadata, when combined with visual data, bolsters model performance. Our experiments have shown that with deliberate training and fine-tuning strategies, multi-modal models can acquire a heightened geographical awareness and exhibit improved classification abilities, even when data is scarce.

Our results indicate that the thoughtful incorporation of available rich metadata in remote sensing datasets can markedly enhance the efficiency of few-shot learning models. The successful fine-tuning of RemoteCLIP to perform geography-aware classification not only confirms the merits of multi-modal learning approaches in the field of remote sensing but also paves the way for further

scholarly inquiry. Subsequent research could investigate the limits of these techniques, refine learning rate optimization, and extend these methods to other complex datasets and scenarios within remote sensing. The encouraging outcomes of this research contribute to the progression of remote sensing technology, ultimately improving our capacity to analyze and comprehend the Earth's topographies and phenomena via sophisticated satellite and aerial imagery.

References

[1] Clifford Broni-Bediako, Yuki Murata, Luiz H. B. Mormille, and Masayasu Atsumi. Searching for cnn architectures for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[2] Si-Bao Chen, Qing-Song Wei, Wen-Zhong Wang, Jin Tang, Bin Luo, and Zu-Yuan Wang. Remote sensing scene classification via multi-branch local attention network. *IEEE Transactions on Image Processing*, 31:99–109, 2022.

[3] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. 2023.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[7] Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[8] Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, and Ho-Jin Choi. Dialogcc: Large-scale multi-modal dialogue dataset. 2022.

[9] Haifeng Li, Zhenqi Cui, Zhiqiang Zhu, Li Chen, Jiawei Zhu, Haozhe Huang, and Chao Tao. Rs-metanet: Deep metametric learning for few-shot remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6983–6994, 2021.

[10] Xiaomin Li, Daqian Shi, Xiaolei Diao, and Hao Xu. Scl-mlnet: Boosting few-shot remote sensing scene classification via self-supervised contrastive learning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2022.

[11] Liangqu Long. Maml-pytorch implementation. https://github.com/dragen1860/MAML-Pytorch, 2018.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2021.

[13] Marc Rußwurm, Sherrie Wang, Marco Körner, and David Lobell. Meta-learning for few-shot land cover classification. 2020.

[14] Kristofer Schlachter, Benjamin Ahlbrand, Zhu Wang, Ken Perlin, and Valerio Ortenzi. Zero-shot multi-modal artist-controlled retrieval and exploration of 3d object sets. 2022.

[15] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. 2017.

[16] Yaw A. Twumasi, Edmund C. Merem, Tomas Ayala-Silva, Albert Osei, Brilliant M. Petja, and Kia Alexander. Techniques of remote sensing and gis as tools for visualizing impact of climate change-induced flood in the southern african region. *American Journal of Climate Change*, 06(02):306–327, 2017.

[17] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. 2017.

[18] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. 2023.

[19] Xin Wang, Lin Duan, Chen Ning, and Huiyu Zhou. Relation-attention networks for remote sensing scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:422–439, 2022.

[20] Yaqing Wang, Quanming Yao, James Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning, 2020.

[21] Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Dino-mc: Self-supervised contrastive learning for remote sensing imagery with multi-sized local crops. 2023.

[22] Lei Xing, Shuai Shao, Yuteng Ma, Yanjiang Wang, Weifeng Liu, and Baodi Liu. Learning to cooperate: Decision fusion method for few-shot remote-sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

[23] Pei Zhang, Ying Li, Dong Wang, and Jiyue Wang. Rs-sskd: Self-supervision equipped with knowledge distillation for few-shot remote sensing scene classification. *Sensors*, 21(5), 2021.

[24] Qin Zou, Lihao Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2321–2325, 2015.

A Task Allocation

Algorithm 2 Data Preparation for Few-Shot Learning

```
df ← LoadData('sorted.csv')
country_to_category ← MapCountriesToCategories(df)
image_counts ← CountImagesPerCountry(df)
set_targets ← ComputeSetTargets(image_counts)
sorted_countries ← SortCountriesByImageCount(image_counts)
sets ← InitializeSets
category_sets ← InitializeCategorySets(df)
PreAllocateCountries(sorted_countries, country_to_category, sets, category_sets)
AllocateRemainingCountries(sorted_countries, country_to_category, sets, set_targets, category_sets)
OutputSetsInfo(sets)
CheckAndPrintCategoryRepresentation(sets, category_sets)
CheckAndPrintSetOverlaps(sets)
SaveSetsToCSV(sets, 'train_set.csv', 'val_set.csv', 'test_set.csv')
```

Algorithm 3 Pre-allocations of Countries (Tasks)

```
function PreAllocateCountries(sorted_countries, country_to_category, sets, category_sets)
    for each category in UniqueCategories(df) do
        for each country in sorted_countries do
            if country_to_category[country] == category then
                if category not in category_sets['train'] then
                    AllocateCountry(country, 'train', sets, category_sets)
                else if category not in category_sets['val'] then
                    AllocateCountry(country, 'val', sets, category_sets)
                else if category not in category_sets['test'] then
                    AllocateCountry(country, 'test', sets, category_sets)
                end if
            end if
        end for
    end for
end function
```

Algorithm 4 Allocation of Remaining Countries (Tasks)

```
function AllocateRemainingCountries(sorted_countries, country_to_category, sets, set_targets, category_sets)
    for each country in sorted_countries not in any set do
        selected_set ← SelectSetBasedOnWeights(sets,
                set_targets, category_sets, country_to_category[country])
        AllocateCountry(country, selected_set, sets, category_sets)
    end for
end function
```

B Supporting Function Definitions

- **LOAD_DATA(file_path)**: Loads data from a CSV file into a DataFrame.
- **MAP_COUNTRIES_TO_CATEGORIES(df)**: Maps each country to its first category in the DataFrame.
- **COUNT_IMAGES_PER_COUNTRY(df)**: Counts images per country, applying a cap and a minimum threshold.
- **COMPUTE_SET_TARGETS(image_counts)**: Calculates target numbers for training, validation, and test sets based on image counts.
- **SORT_COUNTRIES_BY_IMAGE_COUNT(image_counts)**: Sorts countries by image count in ascending order for fair allocation.
- **INITIALIZE_SETS()**: Initializes and returns empty sets for training, validation, and test splits.
- **INITIALIZE_CATEGORY_SETS(df)**: Initializes and returns sets for keeping track of categories in each split.
- **OUTPUT_SETS_INFO(sets)**: Prints information about the distribution of images and countries in each set.
- **CHECK_AND_PRINT_CATEGORY_REPRESENTATION(sets, category_sets)**: Checks and prints whether all categories are represented in each set.
- **CHECK_AND_PRINT_SET_OVERLAPS(sets)**: Checks and prints any overlaps between the training, validation, and test sets.
- **SAVE_SETS_TO_CSV(sets, train_file, val_file, test_file)**: Saves the specified sets to CSV files.

C Training Curves for RemoteCLIP with Semantic Injection

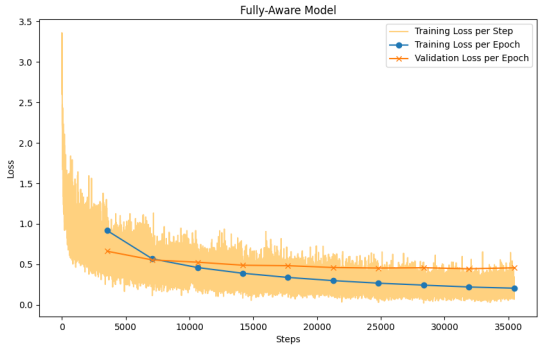


Figure 4: RemoteCLIP Semantic Injection without Dropout

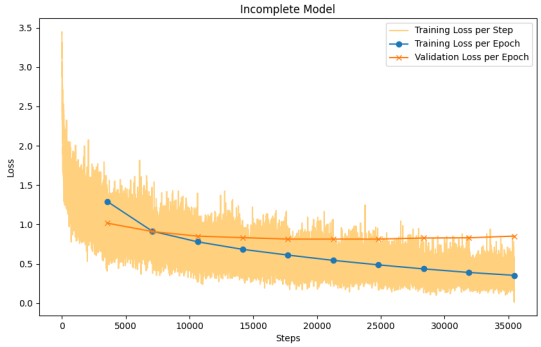


Figure 5: RemoteCLIP Semantic Injection with Dropout

D FMoW Dataset Visualized

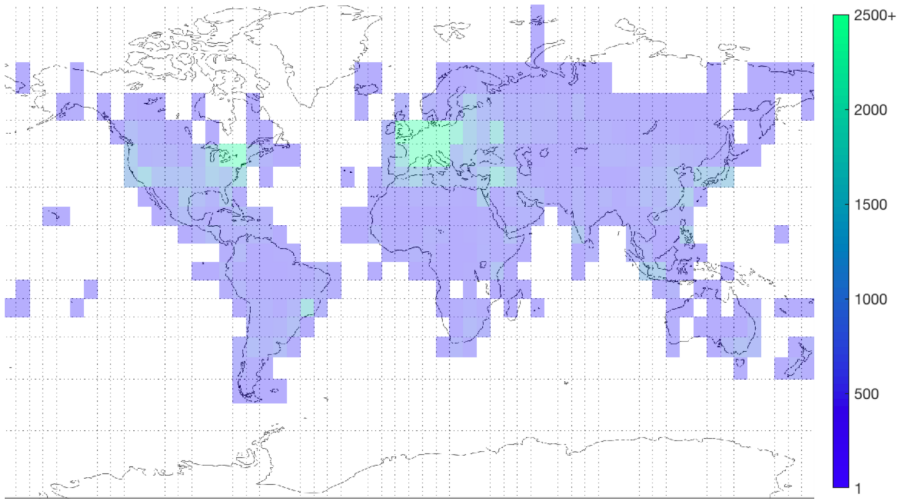


Figure 6: FMoW Image Distribution