# Geographical Priors for Fine-grained Butterfly Species Classification

Navodita Mathur
Independent Researcher
nam266@pitt.edu

Aishwarya Kotharu
Independent Researcher
ak2722@cornell.edu

## Abstract

*Butterflies are key bioindicators, yet their visual classification is confounded by strong geographic and ecological variability. We investigate whether adding habitat context from satellite imagery can improve recognition accuracy. Our pipeline fuses in-situ photographs with Sentinel 2 composites from a 2.5 km area of interest, together with spatiotemporal metadata (location and month), using a ResNet 50 image encoder with Feature wise Linear Modulation (FiLM) for feature fusion. On a 40 species dataset (about 6k observations), models enriched with multispectral and contextual features consistently outperform image-only baselines. The best configuration, which combines multispectral imagery with spatio-temporal data conditioned through FiLM, achieves 92% Top-1 accuracy in the standard test set and 48% on a rare class split, showing that habitat-aware multimodal learning is a promising direction for fine-grained species classification.*

## 1. Introduction

Traditional fine-grained classification models for species identification primarily rely on image-based features derived from in-situ photographs. While effective to a degree, these approaches often overlook the broader ecological context that shapes species distributions and traits. Environmental factors such as vegetation structure, land cover, and climate play a crucial role in determining suitable habitats and influencing phenotypic variability, which in turn help in monitoring and classification. Remote sensing offers high-resolution, multispectral data capable of capturing detailed habitat characteristics at scale. In this study, we investigate the potential of integrating Sentinel-2 imagery with conventional visual inputs along with spatiotemporal data to improve butterfly species classification. By combining species-specific morphological cues with habitat-level context, we aim to enhance classification performance and develop a more ecologically informed, scalable framework for biodiversity monitoring and research.

## 2. Related Work

Fine-grained visual classification (FGVC) has seen significant progress in recent years, particularly in biodiversity applications where distinguishing between visually similar species is critical. Deep learning models, including convolutional neural networks (CNNs) and transformer-based architectures, achieve strong performance on large-scale datasets such as iNaturalist and NABirds, which span thousands of species across taxa [14]. Nevertheless, FGVC remains challenging due to high intraclass variability and subtle interclass distinctions, especially under natural conditions with heterogeneous lighting, complex backgrounds, and varied viewpoints.

To mitigate these challenges, researchers have increasingly moved beyond morphology and incorporated contextual information. Species identity is often tightly constrained by ecological and spatiotemporal factors such as location, seasonality, and habitat. Leveraging this, several studies have demonstrated that metadata, including geographic coordinates, observation time, and habitat descriptors, can significantly improve classification accuracy. For example, Chu et al.[1], Mac[6], and Nguyen et al.[7] show notable gains when augmenting visual features with auxiliary metadata, while Skreta et al.[12] highlight the value of spatiotemporal signals for butterfly classification. These works underline that species recognition is as much about ecological context as it is about morphology.

Remote sensing offers a powerful way to capture such ecological context at scale. Multispectral satellite imagery from platforms such as Sentinel-2 and Landsat encodes detailed information about vegetation, land cover, and environmental gradients that structure species distributions [9] . While this data has long been used in species distribution modeling (SDM) and habitat suitability assessments [3, 4], it has only recently begun to be integrated directly into fine-grained classification. This has motivated new work that combines ground-level observations with habitat-scale remote sensing signals.

Recent multimodal representation learning approaches demonstrate the potential of such integration. BirdSAT [11] jointly learns from ground-level and satellite imagery through cross-view contrastive masked autoencoders, showing that ecological context can sharpen bird classification. TaxaBind [10] introduces a unified embedding space for classification, retrieval, and distribution modeling, emphasizing the scalability of multimodal ecological frameworks. WildSAT [2], in turn, uses citizen science wildlife observations to supervise habitat-sensitive satellite representations by aligning remote sensing data with habitat descriptions, occurrence records, and location encoders. These enriched embeddings support downstream prediction tasks such as bird encounter rate estimation and cross-image retrieval.

Together, these studies highlight an important shift: from treating FGVC as a purely visual problem to embracing ecologically grounded multimodal learning. Much of the recent progress has been driven by large foundation models, with FGVC cast as a downstream task. However, these successes need not hinge on scale alone. Careful alignment between visual, spatial, and ecological modalities can itself yield strong gains even for simpler models and limited data regimes.

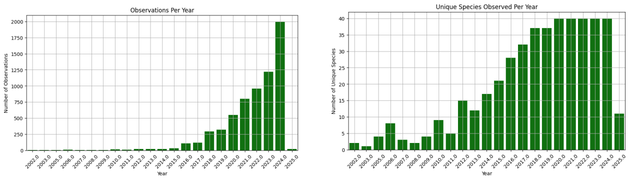
Building on this perspective, our work investigates how Sentinel-2 imagery can be fused with species-specific visual cues to improve butterfly classification. By combining fine-grained appearance features with broader habitat context, we aim to demonstrate that ecologically grounded alignment can drive progress in biodiversity monitoring.

# 3. Methodology

## 3.1. Dataset

To demonstrate this, we constructed a fine-grained butterfly classification dataset by querying the iNaturalist platform for observations from 40 species (≈6000 image-based records). Each observation includes the contributed photograph and structured metadata (GPS coordinates and timestamp), spanning diverse habitats and capture conditions. We also define a *Rare* split comprising species whose training frequency lies below the 20th percentile of per-species counts.

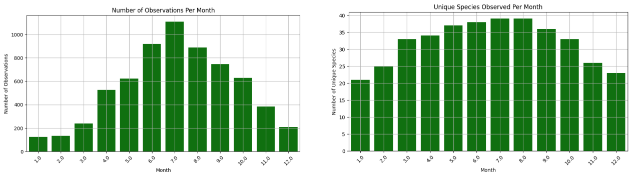
To provide habitat context, for each observation we extract a Sentinel-2 surface reflectance imagery over a **2.5 km** square area-of-interest (AOI) centered at the GPS location within a ±**7**-day window around the timestamp with less than 20% cloud coverage. We use Google Earth Engine to download imagery with 9 bands and compute indices as described in 3.2.3.



(a) Observations by year

(b) Unique species by year

Figure 1. Temporal distribution of observations and species.



(a) Observations by month

(b) Unique species by month

Figure 2. Seasonal distribution of observations and species.

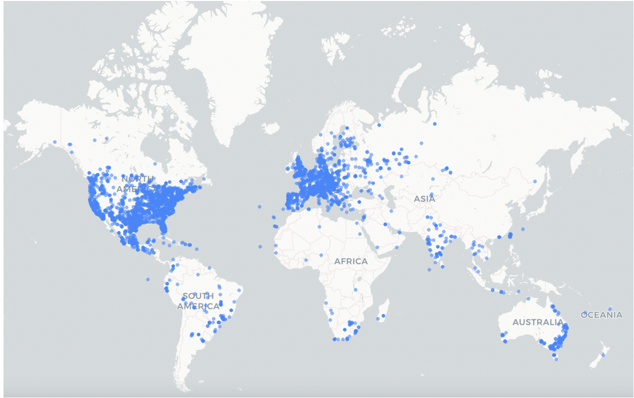


Figure 3. Geographic locations of observations.

The entire dataset was split using 80-10-10 rule and the rare species dataset constructed out of the test split had 56 observations and 8 species.

## 3.2. Architecture

Our approach integrates three sources of information: (1) in-situ butterfly photographs, (2) satellite imagery, and (3) structured metadata consiting of location (lat, lon), time (month and year) and indices (Normalized Difference Vegetation Index, Enhanced Vegetation Index,and Green Chlorophyll Index). Each modality is encoded separately and fused for classification of the species.

### 3.2.1. Image Backbone

We use a ResNet-50 [5] initialized on ImageNet and fine-tuned it on the dataset with Cross-Entropy loss. Images are resized to 224 × 224 and augmented with vertical and horizontal flips, and color jitter.

### 3.2.2. Satellite Imagery backbone

For the **RGB** satellite ablation, we feed a 3-band Sentinel-2 RGB composite to an ImageNet-pretrained ResNet-50 (fine-tuned end to end).

For the *multispectral* ablation, we use a TorchGeo[13] ResNet-50 initialized with Sentinel-2 multi-spectral SATLAS weights, that accepts bands B2, B3, B4, B5, B6, B7, B8, B11 and B12.

### 3.2.3. Metadata & Index Encoding

We encode location (latitude, longitude), time (month, year), and the vegetation indices (NDVI/EVI/GCI) with a two-layer MLP (ReLU + dropout). To respect periodicity, month is mapped via sine/cosine:

$$\sin_m = \sin\left(2\pi \frac{m}{12}\right), \quad \cos_m = \cos\left(2\pi \frac{m}{12}\right)$$

Latitude/longitude are converted to radians and similarly mapped:

$$\sin_{\text{lat}} = \sin(\text{lat}_{\text{rad}}), \quad \cos_{\text{lat}} = \cos(\text{lat}_{\text{rad}}),$$

$$\sin_{\text{lon}} = \sin(\text{lon}_{\text{rad}}), \quad \cos_{\text{lon}} = \cos(\text{lon}_{\text{rad}})$$

We also extracted vegetation indices from the Sentinel-2 bands, specifically:

- NDVI (Normalized Difference Vegetation Index):

$$\text{NDVI} = \frac{NIR - RED}{NIR + RED}$$

- EVI (Enhanced Vegetation Index):

$$\text{EVI} = G \cdot \frac{NIR - RED}{NIR + C_1 \cdot RED - C_2 \cdot BLUE + L}$$

  where $G = 2.5$, $C_1 = 6$, $C_2 = 7.5$, and $L = 1$.
- GCI (Green Chlorophyll Index):

$$\text{GCI} = \frac{NIR}{GREEN} - 1$$

The indices and year were standardized using min-max scaling and the combined vector was fed into an MLP with two hidden layers to generate a fixed-size embedding vector.

### 3.2.4. Fusion and Classification

The image embedding, satellite embedding, and metadata embedding are concatenated and passed to a one-layer MLP classifier with 0.2 dropout, producing a softmax over the 40 species.

| Hyperparameter | Value |
|---|---|
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-6}$ |
| Weight decay | $1 \times 10^{-5}$ |
| Epochs | 100 |
| Batch size | 32 |
| Loss | Cross-Entropy |
| Backbone | ResNet-50 |
| LR scheduler | ReduceLROnPlateau |
| Monitor | Validation loss |
| Mode | `min` |
| Factor | 0.2 |
| Patience | 3 |

Table 1. Training hyperparameters.

### 3.2.5. Feature Modulation (FiLM)

To strengthen cross-modal interaction, we condition intermediate image features on ecological context using *Feature-wise Linear Modulation* (FiLM) [8]. A one-layer MLP maps the metadata embedding to per-channel $\gamma, \beta$ applied to selected ResNet-50 blocks:

$$\text{FiLM}(x) = \gamma \odot x + \beta$$

This encourages the visual pathway to adapt to habitat and season cues (e.g., vegetation and phenology). We ablate FiLM on/off and report detailed results in Table 4.

## 3.3. Training Strategy

We use Adam optimizer. The initial learning rate of $10^{-6}$ is chosen. The `ReduceLROnPlateau` scheduler monitors validation loss and multiplies the current learning rate by 0.2 after 3 epochs without improvement, and a floor at $10^{-8}$.The best checkpoint is selected by validation loss over the 100-epoch run. In most cases the following hyperparameters were used (Table 1).

# 4. Results

Table 2 reports Top-1 accuracy across core training configurations. Table 3 focuses on Feature-wise Linear Modulation (FiLM) with multispectral satellite imagery. The best results in each column are bolded.

## 4.1. Key Findings

- **Satellite cues drive the largest gains**: Adding satellite imagery improves accuracy by up to **11% on the Test split** and **13% on the Rare split**. In contrast, using only geographic and temporal metadata (without imagery) yields more modest improvements of 4. 8% (test) and 4% (rare).
- **Multispectral advantages**: Switching from a frozen to an end-to-end multispectral pipeline raises Rare accuracy

Table 2. Results with configurations consisting of metadata (Location & Time) and RGB satellite imagery (referred to as Sat (RGB)) . Note: Satellite Imagery encoders here are pretrained on ImageNet and fine-tuned on the curated dataset

| Configuration | Test | Rare |
|---|---|---|
| Image only | 0.764 | 0.32 |
| Image + Location + Time | 0.812 | 0.36 |
| Image + Location + Time (FiLM) | 0.880 | 0.41 |
| Image + Sat (RGB) | 0.877 | 0.46 |
| Image + Location + Sat (RGB) | 0.878 | 0.44 |
| Image + Location + Time + Sat (RGB) | 0.880 | 0.44 |
| Image + Location + Sat (RGB; FiLM) | **0.891** | **0.50** |
| Image + Location + Time + Sat (RGB; FiLM) | 0.883 | **0.50** |

Table 3. Results with configurations consisting of FiLM with multispectral (MS; fine tuned) satellite imagery. Note: We omit vegetation indices since multispectral bands already capture this information.

| Configuration | Test | Rare |
|---|---|---|
| Image + Sat | 0.877 | **0.48** |
| Image + Sat + Location(FiLM) | 0.913 | 0.44 |
| Image + Sat + Location+Time(FiLM) | **0.920** | **0.48** |

from 0.38 to **0.48**, outperforming ImageNet-pretrained RGB bands (0.46 Rare).

- **Role of FiLM**: Without FiLM, augmenting MS with Location+Time increases Test accuracy to 0.895 but reduces Rare to 0.42. With FiLM, the same cues push Rare to **0.50**, while keeping the test accuracy competitive (0.880-0.891).
- **Best configurations**: The highest overall Test accuracy is **0.920** (MS + Location+Time with FiLM), while the best Rare accuracy is **0.50** (RGB + FiLM with Location or Location+Time).

Interestingly, we observe that adding multispectral (MS) cues improves overall Test accuracy (from 0.891 with Location and 0.883 with Location+Time using RGB+FiLM to 0.913 and 0.920 with MS+FiLM), but reduces Rare accuracy (from 0.50 with RGB+FiLM to 0.44–0.46 with MS+FiLM). We believe this may be as many rare taxa co-occur with common species in similar habitats, and subtle spectral or structural differences may not be captured by Sentinel-2 imagery and indices in particular, leading FiLM to reinforce habitat-level shortcuts. To address this, we plan to incorporate higher-resolution PlanetScope imagery and leverage pre-trained encoders such as those provided by taxabind[10] and WildSat[2] to learn richer, species-aware spectral features that can improve generalization to rare classes.

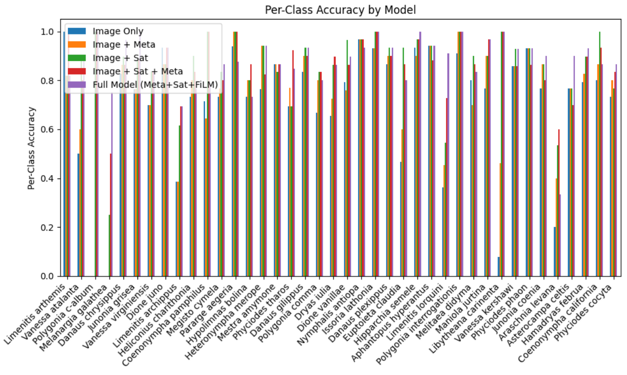


Figure 4. Error analysis. Confusions concentrate among visually similar species; adding Sentinel-2 and FiLM reduces several high-frequency off-diagonal blocks.

## 4.2. Error Analysis

Figure 4 shows confusion patterns. Errors concentrate among visually similar species; adding Sentinel-2 multispectral cues and FiLM reduces several high-frequency off-diagonal blocks.

## 4.3. Summary

Overall, FiLM-based fusion of satellite imagery with geographic and temporal metadata provides a **principled and robust approach**. It seems to consistently outperform image-only baselines and simple concatenation, and improves recognition of rare species, a crucial factor for ecological monitoring and biodiversity applications.

Detailed data are shown in the appendix (FiLM source-by-source in Table 4 and multispectral variants in Table 5) where we see how location and time drives the largest gain in experiments as opposed to using Indices, particularly when taking rare species into account.

# 5. Conclusion

We studied context-aware fine-grained species classification by fusing in-situ photographs with Sentinel-2 habitat cues and structured ecological metadata. Our simple recipe, ResNet-50 for images, ResNet-50 for satellite image and MLP with a linear layer for AOI-aggregated NDVI/EVI/GCI features, and sine/cosine encodings of location and month, with optional FiLM conditioning, consistently improves over image-only baselines.

This work has one main limitation. Evaluation uses random (non-blocked) splits and therefore does not enforce geographic or temporal disjointness.

## 6. Future Work

We see several directions to strengthen both methodology and evaluation:

- **Spatially blocked evaluation:** adopt geographic block cross-validation to prevent train/test location overlap and quantify out-of-region generalization.
- **Temporally blocked evaluation:** year-held-out or forward-chaining splits to assess robustness under seasonal/yearly distribution shift.
- **Long-tailed learning and calibration:** incorporate class-balanced or focal losses, temperature scaling, and selective prediction/abstention to better serve rare species.

## 7. Acknowledgements

Species-occurrence data were obtained from iNaturalist (https://www.inaturalist.org ), contributed by individual observers and shared under Creative Commons licenses; we used only items with compatible CC licenses (excluding All Rights Reserved and NoDerivatives) and provide per-image attribution.

## 8. Appendix

In tables 4 and 5, FM denotes Feature Modulation (FiLM). Sat denotes Sentinel-2 imagery downloaded from Google Earth Engine encoded using ResNet-50. “Frozen” keeps the satellite image encoder backbone fixed; “Fine-tuned” updates entire encoder.

**Takeaways for Table 4.**

- Fine-tuning generally outperforms frozen backbones when fusing satellite cues.
- Location and month metadata help, and FM further boosts performance by conditioning the image encoder on context.
- Without Satellite Imagery, FiLM recovers a portion of the gains; with Satellite Imagery, FiLM yields the strongest accuracy on both Test and Rare splits.

**Takeaways for Table 5.**

- Multispectral fusion improves over image-only baselines; end-to-end fine-tuning show significant improvements over frozen encoder.
- The best configuration combines Location+Time metadata, Sentinel-2 features, and FM, reaching the top Test and Rare scores.

**Link to main results.** These detailed ablations support the headline results in Table 2, showing consistent gains from Sentinel-2 fusion and additional improvements from FiLM.

Table 4. Top-1 accuracy on the standard Test set and the Rare Test set for different input configurations. FM-Feature Modulation, Sat-RGB Satellite Imagery at 10m resolution

| **Configuration** | **Test** | **Rare** |
|---|---|---|
| *Baseline Models* | | |
| Image Only | 0.764 | 0.32 |
| Image + Sat | 0.877 | 0.46 |
| Image + Meta | 0.8014 | 0.32 |
| Image + Meta + Sat | 0.873 | 0.48 |
| *Feature Modulation Enabled* | | |
| Image + Meta | 0.864 | 0.40 |
| Image + Meta + Sat | 0.884 | 0.48 |
| *Without FM, Without Sat* | | |
| Location Only | 0.8041 | 0.34 |
| Location + Time | 0.812 | 0.36 |
| Indices Only | 0.792 | 0.32 |
| Without Year | 0.805 | 0.38 |
| *Without FM, With Sat* | | |
| Location Only | 0.876 | 0.44 |
| Location + Time | 0.88 | 0.44 |
| Indices Only | 0.883 | 0.42 |
| Without Year | 0.875 | 0.46 |
| *With FM Without Sat* | | |
| Location Only | 0.867 | 0.46 |
| Location + Time | 0.880 | 0.42 |
| Indices Only | 0.827 | 0.30 |
| Without Year | 0.849 | 0.38 |
| *With FM With Sat* | | |
| Location Only | 0.891 | 0.50 |
| Location + Time | 0.883 | 0.50 |
| Indices Only | 0.865 | 0.40 |
| Without Year | 0.899 | 0.48 |

## References

[1] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019. 1

[2] Rangel Daroya, Elijah Cole, Oisin Mac Aodha, Grant Van Horn, and Subhransu Maji. Wildsat: Learning satellite image representations from wildlife observations, 2025. 2, 4

[3] Johannes Dollinger, Philipp Brun, Vivien Sainte Fare Garnot, and Jan Wegner. Sat-sinr: High-resolution species distribution models through satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-2-2024:41–48, 2024. 1

[4] Lauren E. Gillespie, Megan Ruffley, and Moises Exposito-Alonso. Deep learning models map rapid plant species changes from citizen science and remote sensing data. *Proceedings of the National Academy of Sciences*, 121(37): e2318296121, 2024. 1

Table 5. Top-1 accuracy on the standard *Test* and *Rare* splits. **Sat Enc (MS)** denotes a multispectral *satellite image encoder* (e.g., Sentinel-2 bands) fused with the photo (RGB) encoder; no handcrafted indices are used. **Frozen** trains only the fusion/classifier layers with encoders fixed; **Fine-tuned** updates the full encoders. FiLM = feature-wise linear modulation.

| **Setting** | **Tuning** | **Test** | **Rare** |
|---|---|---|---|
| *Image + Sat Enc (MS) fused* | | | |
| Image + Sat Enc (MS) | Frozen | 0.838 | 0.380 |
| Image + Sat Enc (MS) | Fine-tuned | 0.877 | **0.480** |
| *Metadata (no FiLM)* | | | |
| Location only | Frozen | 0.800 | 0.380 |
| Location only | Fine-tuned | 0.892 | 0.440 |
| Location + Time | Frozen | 0.820 | 0.360 |
| Location + Time | Fine-tuned | 0.895 | 0.420 |
| *Metadata (with FiLM)* | | | |
| Location only | Frozen | 0.875 | 0.440 |
| Location only | Fine-tuned | 0.913 | 0.440 |
| Location + Time | Frozen | 0.879 | 0.460 |
| Location + Time | Fine-tuned | **0.920** | **0.480** |

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 2

[6] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019. 1

[7] Tin Nguyen, Peijie Chen, and Anh Totti Nguyen. Leveraging habitat information for fine-grained bird identification, 2025. 1

[8] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. 3

[9] Nathalie Pettorelli, William F. Laurance, Timothy G. O'Brien, Martin Wegmann, Harini Nagendra, and Woody Turner. Satellite remote sensing for applied ecologists: opportunities and challenges. *Journal of Applied Ecology*, 51 (4):839–848, 2014. 1

[10] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Adeel Ahmad, and Nathan Jacobs. Taxabind: A unified embedding space for ecological applications, 2024. 2, 4

[11] Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, and Nathan Jacobs. Birdsat: Cross-view contrastive masked autoencoders for bird species classification and mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7136–7145, 2024. 2

[12] Marta Skreta, Sasha Luccioni, and David Rolnick. Spatiotemporal features improve fine-grained butterfly image classification. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*, 2020. 1

[13] Adam J. Stewart, Caleb Robinson, Isaac A. Corley, Anthony Ortiz, Juan M. Lavista Ferres, and Arindam Banerjee. Torchgeo: Deep learning with geospatial data, 2022. 3

[14] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1