BirdSAT: Cross-View Contrastive Masked Autoencoders for Bird Species Classification and Mapping

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, Nathan Jacobs
Washington University in St. Louis
{s.sastry, k.subash, a.dhakal, di.huang, jacobsn}@wustl.edu

Abstract

We propose a metadata-aware self-supervised learning (SSL) framework useful for fine-grained classification and ecological mapping of bird species around the world. Our framework unifies two SSL strategies: Contrastive Learning (CL) and Masked Image Modeling (MIM), while also enriching the embedding space with metadata available with ground-level imagery of birds. We separately train uni-modal and cross-modal ViT on a novel cross-view global bird species dataset containing ground-level imagery, metadata (location, time), and corresponding satellite imagery. We demonstrate that our models learn fine-grained and geographically conditioned features of birds, by evaluating on two downstream tasks: fine-grained visual classification (FGVC) and cross-modal retrieval. Pre-trained models learned using our framework achieve SotA performance on FGVC of iNAT-2021 birds and in transfer learning settings for CUB-200-2011 and NABirds datasets. Moreover, the impressive cross-modal retrieval performance of our model enables the creation of species distribution maps across any geographic region. The dataset and source code will be released at `https://github.com/mvrl/BirdSAT`.

1. Introduction

Species classification and distribution mapping are two important tasks for ecologists who monitor and protect the habitats of endangered species. Species classification involves categorizing species with subtle differences into fine-grained classes. It lies within the continuous manifold of basic visual recognition tasks and more complex visual identification tasks. Moreover, it coincides with the task of Fine-Grained Visual Classification (FGVC) which has already been used for distinguishing between models of cars [1], species of birds [2, 3, 4], airplanes [5], etc. On the other hand, the task of species distribution mapping aims at mapping the habitation of species of interest over any geographic region in the world. In this work, we propose to learn a unified representation space useful for solving both of these tasks. Specifically, we evaluate our framework for classifying and mapping `bird` species around the world. However, our models are general enough to be easily extended to `any` species of interest.

As easy as it may sound, low inter-class variance and high intra-class variance make the task of species classification relatively difficult. Most often, species in the same category vary in terms of their pose, size, and lighting. This makes it challenging for deep learning models to extract category-specific rich features useful for fine-grained classification and mapping. Previous works have approached this challenge in one of the following ways: (1) Collecting additional labeled data [6]; (2) Using sophisticated learning techniques [7, 8]; (3) Using auxiliary and/or metadata as additional cue [9, 10]. Out of these, (1) is usually the most time-consuming and expensive approach and (2) requires careful design of objectives and methods for effective results. However, the inclusion of metadata has proven to be very effective.

The task of species classification requires fine-grained visual representation learning capabilities. Recently, self-supervised learning (SSL) strategies such as Masked Image Modeling (MIM) [11, 12] have proven to be useful for learning discriminative features. On the other hand, the task of species mapping, which can be realized as a cross-modal retrieval task, would benefit from contrastive learning (CL) based SSL. To learn a common embedding space for both of these tasks, we propose to use a general SSL framework trained using objective functions for both MIM and CL.

We expect geolocation and time to provide useful cues for species classification and mapping. Therefore, we incorporate metadata (location, time) into our SSL framework, as additional information to learn from. However, metadata alone is not sufficient for species mapping as noted by other works on geo-aware mapping tasks [13, 14]. Therefore, we additionally incorporate cross-view visual information by collecting freely available corresponding satellite images for each ground-level image. We expect that these images

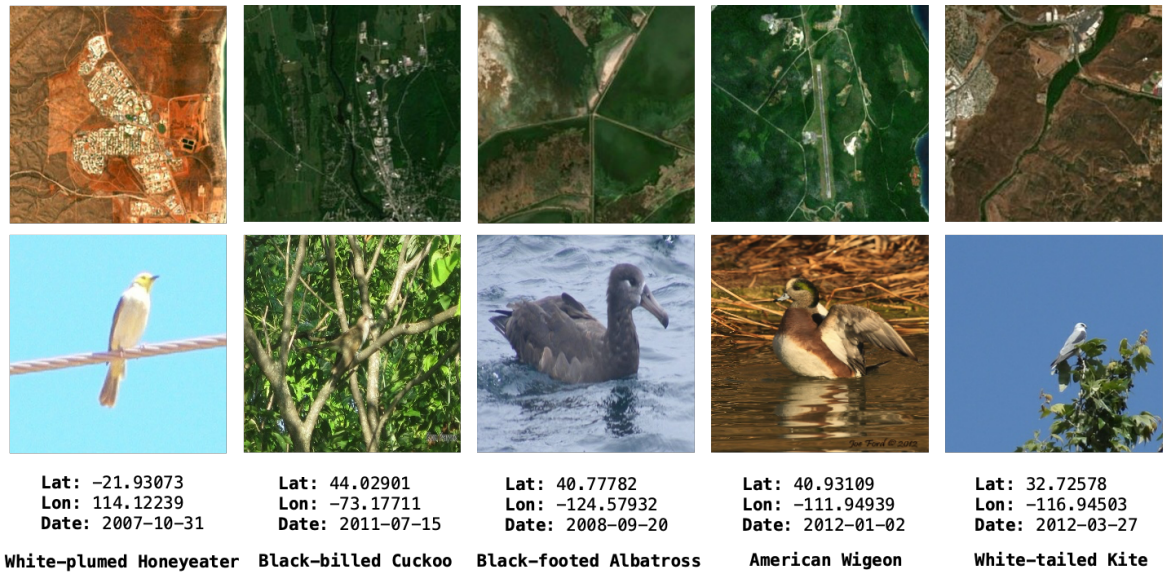


Figure 1: **Cross-View iNAT-2021 Birds Dataset**. Examples of paired satellite and ground-level images of birds along with metadata associated with each pair.

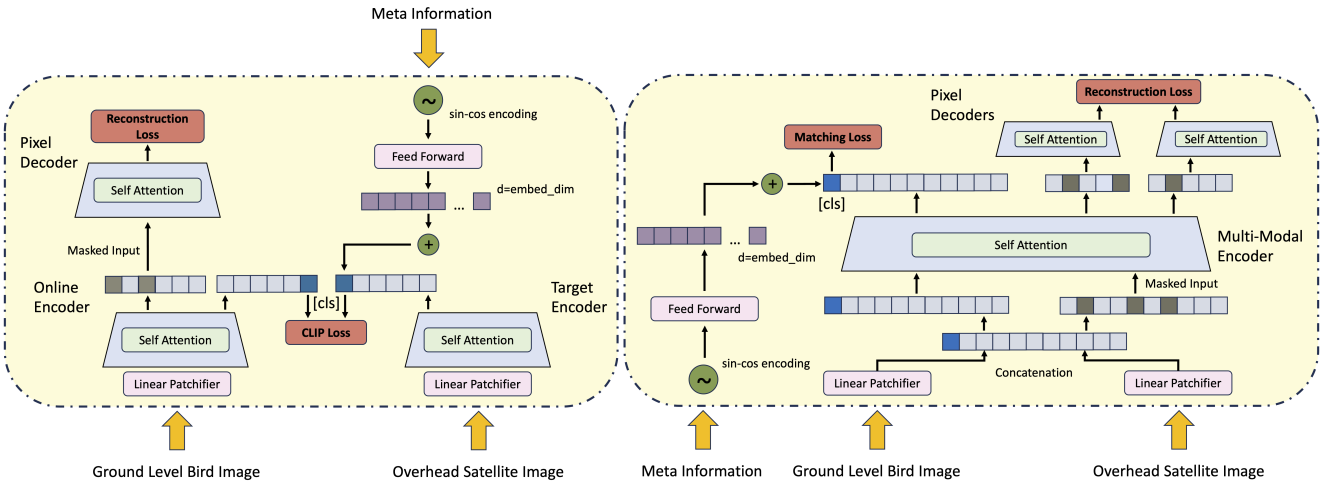
shall provide the model with a context of the surroundings and habitat a bird might be found. Further, information coming from multiple modalities is usually combined using early-fusion or late-fusion style architectures. Early-fusion style architectures [15, 16] use a multi-modal model to encode all input information while the latter uses modality-specific models to learn correlated information between the modalities [17, 18]. In this work, we explore and evaluate these kinds of architectures from heuristic and systematic perspectives. In the end, using the models, we are able to map species of birds across the globe at a fine-grained level. The contributions of our work are threefold:

- We introduce a global Cross-View iNAT 2021 Birds Dataset, which contains paired satellite images and corresponding ground-level bird images.
- We propose a framework for cross-view pre-training of vision transformers along with metadata enabling the ecological mapping of bird species.
- We demonstrate the rich representational capability of our pre-trained models by demonstrating SotA on fine-grained bird classification across three datasets.

2. Related Work

Self-Supervised Learning (SSL) has proven to be an effective pre-training strategy for various downstream tasks in computer vision. The two most successful SSL strategies are: Contrastive Learning (CL) and Masked Image Modeling (MIM). Contrastive learning-based pre-training pulls positive pair of samples closer while pushing the negative pairs farther in the embedding space. CL creates a representation space with high instance discriminability useful for various visual recognition tasks. On the other hand, inspired by Masked Language Modeling (MLM) [19] in Natural Language Processing (NLP), MIM has proven to be an effective pre-training strategy in computer vision. MIM is effective, especially for tasks where learning fine-grained concepts is important. MIM was first introduced for vision tasks by Masked Autoencoder (MAE) [11]. MAE has since been used as a representation learning framework for video [20], as well as for other visual modalities such as satellite imagery [12, 21]. Moreover, owing to its flexibility and scalability, MAE has also been adapted for different multi-modal representation learning frameworks such as MultiMAE [15] and M3AE [22]. MAE frameworks offer rich representational capability useful for fine-grained tasks, however, the limited discriminability of its embedding space hampers the performance on visual recognition tasks. To mitigate this, some of the recent works [23, 24, 25] have proposed to introduce CL-style learning into the MIM-based MAE framework. One of the main downstream tasks of our work is fine-grained visual recognition. Therefore, we also pre-train our proposed cross-view framework using both contrastive and MIM losses.

Fine-Grained Visual Classification (FGVC) requires distinguishing subtle yet discriminative details within a category (e.g., animal, bird, car, etc.). Accordingly, most of the prior works have proposed different attention mechanisms [26, 27, 28, 29] which detect the discriminative



(a) Cross-View Embed MAE

(b) Cross-View Metric MAE

Figure 2: **Our proposed framework**. We evaluate (a) uni-modal (late-fusion) and (b) cross-modal (early-fusion) pre-training of ViT incorporating metadata and contrastive and masked reconstruction objectives.

parts in an image and amplify their corresponding features for recognition. Moreover, in order to enhance fine-grained representations, different modules that are easy to be plugged into common backbone architectures have been proposed [8, 30]. In a separate line of work, different SSL techniques have been introduced either as an additional self-supervision [31] or as a pre-training strategy for FGVC. For example, Yu et.al. [31] propose randomly masking parts of an image and forcing the network to predict the position of the masked parts. Different pre-text tasks for SSL such as jigsaw solving, adversarial learning, and SimCLR [32] based CL are explored in [33]. In [34], a multi-stage SSL strategy is proposed, where a SimCLR-style framework is trained with images progressively degraded with masks having different granularity at each stage. In a recent work [35] an additional GradCAM-guided loss is introduced into a MoCo-style SSL framework. In our work, inspired by the success of SSL in various computer vision tasks including FGVC, we propose a cross-view SSL framework trained on our novel dataset containing ground-level images paired with their corresponding satellite imagery.

Geography-Aware Learning leverages the high-level context available in the geolocation of any ground-level scene. Such information proves to be a valuable signal for various visual recognition tasks [36, 37, 38, 39] and has been successfully used for mapping the distribution of different attributes across a geographic region of interest [40, 41, 13, 14]. For example, Tang et.al. [36] proposed to encode location information into their network yielding improved visual recognition performance. Ayush et.al. [42] proposed adding a geolocation classification loss into the original MoCo-v2 SSL framework achieving performance gain in a diverse range of remote sensing tasks. Similarly, from some of the recent works [38, 37], it has become evident that including geographic information improves the performance on the task of FGVC. Inspired by these findings, in both of the cross-view SSL frameworks proposed in our work, we encode location and date as extra metadata that the model can learn from.

3. Cross-View iNAT-2021 Birds Dataset

We construct a cross-view birds dataset that consists of paired ground-level bird images and satellite images as shown in Figure 1. We expect that this kind of dataset will not only help improve the performance of existing methods but also enable innovative new methods for bird distribution modeling. To do this, we select the iNAT-2021 dataset [2] which spans all over the globe. This dataset is both large scale and contains rich metadata such as geolocation and timestamp of an image.

We carefully filter images of bird species from the dataset which have geolocation associated with them. This resulted in dropping only 888 out of 414,847 (0.2%) observations in training. In testing, we dropped 29 out of 14,860 (0.1%). This did not significantly impact the distribution of the classes (more details in Appendix Section A). Using the geolocation information, we collect Sentinel-2 level 2A images corresponding to each of the ground-level bird images. Each Sentinel-2 image we extract is of resolution 256x256 which spans an area of 6.55 km^2 on the Earth's surface. In total, the dataset contains 413,959 pairs for training and 14,831 pairs for testing.

4. Method

We employ and evaluate two different approaches for contrastively training MAE with satellite images and ground-level bird images. We describe the two approaches in the following sections.

4.1. Cross-View Embed MAE

The overall framework of Cross-View Embed MAE (CVE-MAE) is illustrated in Figure 2 (a). Our method consists of two separate modality-specific transformer encoders and a single transformer decoder for reconstructing the ground-level image modality. Both the encoders have the same architecture based on ViTAE [43], while the decoder has the same architecture as employed by the authors in MAE. The satellite image encoder is directly taken from [44] and is kept frozen throughout.

Similar to other contrastive learning frameworks [32, 17], the ground-level image encoder serves as the online encoder and the satellite image encoder serves as the target encoder. Both the encoders have an extra [cls] token which we use for computing a contrastive objective. The contrastive objective we use in this study is the symmetric InfoNCE loss as used in CLIP [18]. If I^g denotes ground-level image and I^s denotes satellite image, CLIP loss is defined by:

$$L_g = -log\frac{exp(I^g \cdot I^s)^+}{exp(I^g \cdot I^s)^+ + \sum_{j=1}^{N-1} exp(I^g \cdot I^s_j)^-} \quad (1)$$

$$L_s = -log\frac{exp(I^s \cdot I^g)^+}{exp(I^s \cdot I^g)^+ + \sum_{j=1}^{N-1} exp(I^s \cdot I^g_j)^-} \quad (2)$$

$$L_c = \frac{L_g + L_s}{2} \quad (3)$$

Here, $I^g \cdot I^s$ is the normalized cosine similarity between the *[cls]* token obtained from ground-level and satellite image encoders respectively. The sum is over a batch of samples and $+$ and $-$ denote positive and negative pairs within the batch respectively. Similar to MAE, the decoder is used to reconstruct ground-level images using masked versions of tokens obtained from the online encoder. A second forward pass is required to train for this objective as the ground-level encoder requires only the unmasked tokens as input. For the reconstruction objective, we use the L_2 loss defined by:

$$L_r = \sum_{j=1}^{N} |\hat{I^g_j} - I^g_j|^2_2 \quad (4)$$

The overall loss is then defined as -

$$L = L_c + L_r \quad (5)$$

Different from the existing method (i.e. CMAE [23]), our method gets rid of the feature decoder layer. We found that training without the feature decoder layer results in stable loss curves.

4.2. Cross-View Metric MAE

The overall framework of Cross-View Metric MAE (CVM-MAE) is illustrated in Figure 2 (b). This kind of training strategy requires a single multi-modal transformer encoder and separate modality-specific transformer decoders. This is a similar setup as used by previous works on multi-modal MAE [22, 16]. The encoder and decoders have the same architecture as employed by the authors in MAE.

The proposed framework starts by concatenating the tokens computed from the ground level and satellite images using separate linear patchifier layers. A [cls] token is appended to the tokens which is later used for computing the matching loss. **Ground-satellite Matching** predicts whether a pair of satellite and ground-level bird images is positive or negative. A single feed-forward layer is appended so as to train for the objective. The matching loss is simply defined as the binary cross entropy loss between ground-truth labels and the output of the feed-forward layer as follows:

$$L_m = \frac{-1}{2N}\sum_{j=1}^{2N}(y_j log(\hat{y_j}) + (1 - y_j)log(1 - \hat{y_j})) \quad (6)$$

Positive and negative pairs of ground-level and satellite images are defined using the batch currently in training. The satellite image batch is simply rolled to create the set of negative pairs. The output [cls] token from the multi-modal encoder is used for computing this matching loss. Intuitively, this token should capture the joint representation of the image pair.

Additionally, a second forward pass is required for computing the reconstruction objectives. This is necessary as the pixel decoders require a masked version of the encoder outputs. Unmasked tokens are first generated at the input stage using the patchifier layers. They are concatenated and sent to the multi-modal encoder. The tokens computed by the multi-modal encoder are separated back to their respective modality. Finally, the tokens (after concatenating with [mask] tokens) are sent to their modality-specific decoders for reconstruction. Again, we use the L_2 loss to train for the reconstruction objective. The overall loss is defined as -

$$L = L_m + L_r \quad (7)$$

4.3. Incorporating Acquisition Metadata

While often ignored, acquisition metadata, such as when and where an image was captured, provides additional context which can improve our ability to interpret the content of an image. It can not only help reduce the number of possible classes but also help improve the interpretability of a model. Our Cross-View iNAT-2021 Birds Dataset provides geolocation and timestamp for each image. In our implementation, we use latitude, longitude, and month attributes

Table 1: Comparison of accuracy (%) achieved by our proposed models and SotA approaches on the standard test set of the iNAT-2021 Birds dataset. We report linear probing (lin) and fine-tuning (ft) accuracy.

Method	Location	Date	Pre-training	#param. (trainable)	#FLOPS	lin	ft
MoCo-V2-Geo [42]	✓	✗	InfoNCE+Geo-Clf.	115M	13.46G	52.44	85.07
MAE [11]	✗	✗	Recons. Loss	115M	13.46G	41.10	83.14
MetaFormer-2 [9]	✓	✓	ImageNet Clf.	81M	16.90G	-	85.34
CVE-MAE	✗	✗	InfoNCE+Recons. Loss	117M	13.46G	38.86	83.78
CVE-MAE-Meta	✓	✓	InfoNCE+Recons. Loss	117M	13.46G	59.26	86.23
CVM-MAE	✗	✗	Matching+Recons. Loss	115M	31.59G	44.25	85.89
CVM-MAE-Meta	✓	✓	Matching+Recons. Loss	115M	31.59G	**63.33**	**87.46**

of the metadata. As each of the attributes is a real number, we encode them using the sin-cos encoding method. They are then passed to a feed-forward layer which outputs an embedding of the same dimension as our single-stream and dual-stream encoders. Finally, they are added to the [cls] token embedding that results from the encoders. We call these models CVE-MAE-Meta and CVM-MAE-Meta. Note that for our dual stream approach, we only add metadata to the *[cls]* token resulting from the satellite image encoder.

4.4. Meta-Dropout

During initial training runs for pre-training and fine-tuning, we noticed a heavy dependence of our models on metadata for minimizing target objectives (Appendix Section C). For several ground-level images of birds, our models seemed to ignore visual information completely. To address this issue, we randomly dropped metadata (25% of the time) during training. Given the flexibility of our models, it is easy to forward the raw features without having the need to add metadata. Another benefit of this strategy is that it improves inference on unseen examples where metadata is not available.

5. Experiments

We evaluate the performance of our proposed models by first pre-training on the Cross-View iNAT-2021 Birds Dataset and then applying them to various tasks. In the following sections, we describe our implementation details, FGVC performance on iNAT-2021 birds, satellite image to ground level image retrieval performance, and transfer learning performance on CUB-200-2011 [4] and NABirds [3].

5.1. Implementation Details

We randomly crop the ground-level images to a resolution of 384x384 and satellite images to a resolution of 224x224. We use the ViT-B/32 and ViT-B/16 architecture for the ground-level images and satellite images respectively. For pre-training, we use the AdamW [45] optimizer with a weight decay of 0.01. We use a learning rate of $1e^{-4}$ along with cosine annealing warm restarts [46]. We also apply TrivialAugment [47].

For linear probing and fine-tuning, we use the AdamW optimizer with a weight decay of 1e-4 for linear probing and 0.2 for fine-tuning. The learning rates are set to 0.1 and $5e^{-5}$ for linear probing and fine-tuning respectively. For fine-tuning, we additionally apply RandAugment [48], mixup [49], CutMix [50] and LabelSmoothing [51].

We use a batch size of 308 across 4 NVIDIA A100 GPUs for all the experiments. Additional details about all our implementations are present in the Appendix.

5.2. Cross-View iNAT-2021 Birds Experiments

After pre-training, we do supervised training on the Cross-View iNAT-2021 Birds Dataset to evaluate the representations learned by our proposed models. This is done by reporting linear probing as well as fine-tuning accuracy scores on the standard test set of the Cross-View iNAT-2021 Birds Dataset. For CVE-MAE-Meta and CVM-MAE-Meta models, we add metadata-dependent features to the *[cls]* token before classification with the linear head. Again, we use a dropout of 0.25 for the metadata. We compare the performance of our models with the following baselines: MoCo-V2-Geo[1] [42], MAE [11] and Metaformer-2 [9].

Results illustrated in Table 1 show that our single stream CVM-MAE-Meta model beats all other models. The incorporation of satellite images during supervised training has helped the metric-based models beat the embedding-based models. For both training strategies, metadata has improved the testing accuracies by at least 1.57%. Notice that there is a large gap in accuracies between linear probing and fine-tuning. This suggests that complete fine-grained knowledge

[1]Please note that we implement our version of cross-view training of MoCo-V2-Geo.

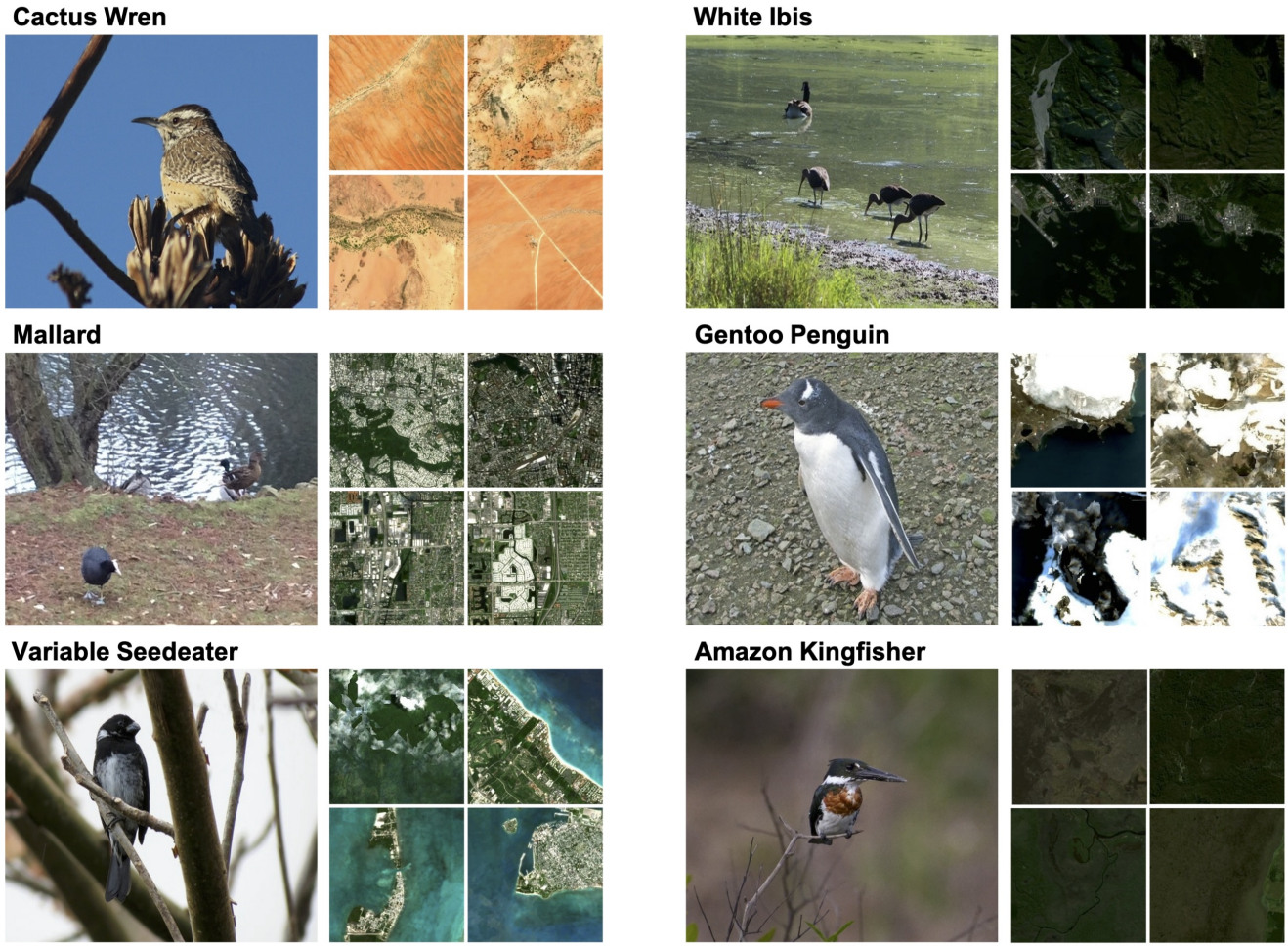


Figure 3: **Top-4 Retrieved Candidates**. We select six different bird species and retrieve the top-4 most similar satellite images (right) to the corresponding bird images (left) in the test set. Notice that the retrieved images are similar to each other while also being relevant for the corresponding bird species.

about the bird species has not yet been embedded into the models during the pre-training stage.

5.3. Zero-Shot Retrieval

The zero-shot retrieval experiment contains two sub-tasks: satellite image to ground-level bird image retrieval and ground-level bird image to satellite image retrieval. We use the former task for computing retrieval metrics while the latter task is for generating species distribution maps. Both retrieval tasks are evaluated using pre-trained models before fine-tuning.

For this first task, we evaluate two different retrieval approaches: 1) Single-stage uni-modal retrieval and 2) Hierarchical cross-modal retrieval. For the single-stage uni-modal retrieval approach, we first compute the [cls] embeddings for all the ground-level bird images and the query satellite image in the test set. Then, we compute the pairwise similarity between the query satellite image embedding and the ground-level image embeddings. Finally, we select the top-k ground-level images with the highest similarity. Additionally, metadata is added for retrieval using our CVE-MAE-Meta model. For calculating the retrieval metrics, we only consider retrieving the correct species rather than the exact image.

The hierarchical cross-modal retrieval approach consists of two stages: 1) selecting candidate ground-level images from similarity computed using the uni-modal model and 2) selecting final ground-level images from the candidates using the matching score computed by the cross-modal model. We propose this approach since embeddings cannot be pre-computed for the cross-modal models and are computationally infeasible for retrieval.

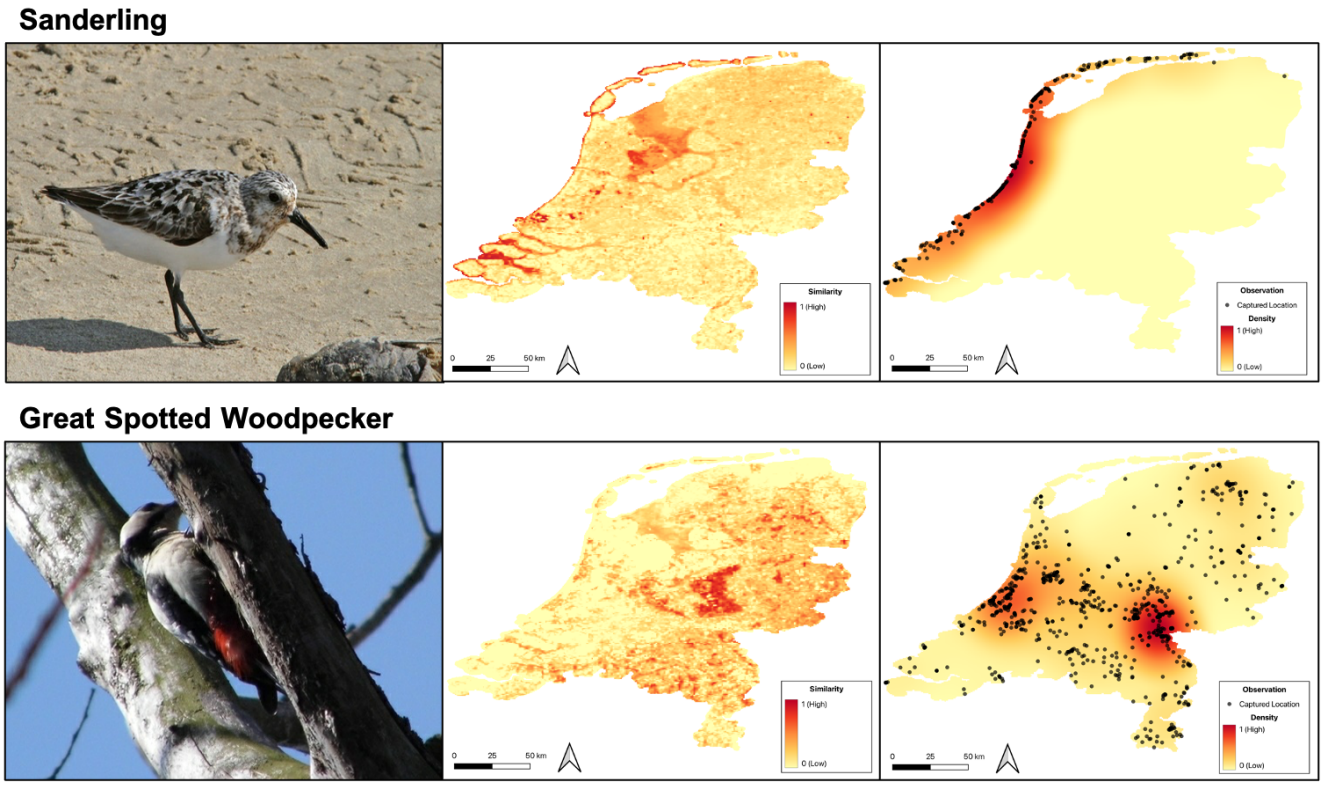


Figure 4: **Generated Bird Distribution Maps**. Using ground-level bird image to satellite image similarity scores, we generate *expected* bird distribution maps over The Netherlands. We show distribution maps of two different species of birds (left) found in The Netherlands taken from the test set of Cross-View iNAT-2021 Birds Dataset (middle). We also show a heatmap of the presence of those species (right) in the iNAT-2021 Birds Dataset.

We show the recall scores (R@5 and R@10) of satellite image to ground-level bird image retrieval in Table 2. The results indicate that satellite images are able to provide a strong cue indicating that habitat and surroundings are important for the retrieval of bird species. On the other hand, both our dual stream models are able to beat MoCo-V2-Geo model indicating that our models have learned more robust embedding spaces. The hierarchical retrieval approach using CVM-MAE-Meta model also performs reasonably well showcasing the effectiveness of first reducing the search space using uni-modal models and then selecting the final candidates using cross-modal models. However, this approach requires pre-training two separate models and searching steps increasing the computational complexity of the overall setup.

Figure 3 depicts examples of satellite images retrieved corresponding to query bird images. Clearly, the retrieved images correspond well with the bird's expected habitat. Further, we generate a species distribution map for two distinct species over The Netherlands (Figure 4). We first collected satellite images over a dense grid draped over The Netherlands. We then interpolated and plotted the ground-level image to satellite image similarity scores. We removed all the observations with similarity scores below zero (more details in Appendix Section D).

To study the true performance of our model, we used bird images from the test set of the Cross-View iNAT-2021 Birds Dataset. We conducted a qualitative evaluation of the maps using the observations present in the ground truth. Vi-

Table 2: Zero-shot satellite image to ground level bird image retrieval results on the standard test set of Cross-View iNAT-2021 Birds Dataset.

Method	R (@5)	R (@10)	#FLOPS
MoCo-V2-Geo [42]	5.77	14.28	29.05G
CVE-MAE	14.45	25.62	29.05G
CVE-MAE-Meta	13.72	26.97	29.05G
CVM-MAE-Meta	**14.52**	**28.88**	60.64G

sually, the maps nicely delineate the presence of birds in the region. Since crowd-sourced datasets are biased towards areas observing high human traffic, the ground truth is not an exhaustive representation of the species distribution. On the other hand, our model is able to provide fine-grained presence of bird species across large geographic regions.

5.4. Transfer Learning Experiments

We evaluate the performance of transfer learning of our model on downstream FGVC Bird datasets. Since satellite images are not available for the majority of the open-sourced datasets, we only study fine-tuning our CVE-MAE-Meta model. We take the best performing CVE-MAE-Meta model on fine-grained classification of iNAT-2021 Birds and then fine-tune it on downstream datasets. We additionally drop metadata during training and inference as the datasets do not include metadata. We consider two of the most popular fine-grained bird classification datasets: CUB-200-2011 and NABirds. The relative scale of these datasets as compared to the iNAT-2021 Birds dataset is presented in Table 3.

Table 3: Various datasets considered in this study.

Dataset	#Training	#Testing	Categories
iNAT-2021 Birds	414,847	14,860	1486
CUB-200-2011	5,994	5,794	200
NABirds	23,929	24,633	555

Table 4: Comparison of accuracy (%) achieved by CVE-MAE-Meta and SotA approaches on the standard test set of CUB-200-211 and NABirds datasets. We report linear probing (lin) and fine-tuning (ft) accuracy.

Method	CUB		NABirds	
	lin	ft	lin	ft
MoCo-V2-Geo [52]	81.19	86.91	83.22	89.26
MAE [11]	80.33	88.46	82.11	89.23
MetaFormer-2 [9]	-	92.40	-	92.70
HERBS [30]	-	93.10	-	93.00
CVE-MAE-Meta	**82.98**	**93.23**	**84.21**	**93.47**

Except for HERBS, all the models were first fine-tuned on iNat-2021 birds dataset. The results in Table 4 show that our model outperforms all other transformer-based models including MoCo-V2-Geo, MAE, Metaformer, and HERBS. The results are consistent for both linear probing and fine-tuning. Our model achieves noticeably high accuracy when linear probing. This indicates that features learned from fine-tuning our pre-trained model on iNAT-2021 Birds are highly robust and transferable.

6. Discussion and Conclusion

In this study, we focused on unifying the problem of fine-grained visual classification (FGVC) and mapping of bird species around the world. We constructed a cross-view dataset consisting of paired ground-level bird images and satellite images. For applications involving ecological mapping and identification of species, satellite images provide spatially correlated topographical information. Therefore, leveraging freely available satellite imagery in SSL enables us to create species maps for any geographic region. Such maps, when augmented with expert knowledge, may be used to refine existing species distribution maps, which otherwise are usually sparse and inaccurate.

We evaluated two architectural frameworks: uni-modal and cross-modal, trained with masked reconstruction and contrastive learning objectives. They differ in the way multi-modal information is fused, which is an essential component for real-time global-scale applications. Uni-modal setup is useful because the modality-specific encoders can be used to pre-compute embeddings. These embeddings can then be used in real-time for a variety of downstream tasks. This becomes essential in large-scale applications. Still, the uni-modal setup is only able to preserve correlated information between the modalities, limiting its performance on recognition tasks. In contrast, the cross-modal setup allows models to learn complementary information coming from a variety of modalities creating rich features useful for complex visual identification tasks. However, the computational complexity of this setup prevents its adoption for large-scale applications. Yet, we conclude that both our training setups are effective at learning general-purpose features that can used for species classification and mapping. Finally, we also presented a two-stage retrieval approach that takes advantage of both the training setups to reduce computation bottleneck.

Owing to the flexibility of our framework, we can easily incorporate other modalities such as text and sound as part of future work. However, one needs to be careful when including several modalities since data collected from uncurated may fail to provide fine-grained information useful for the task of FGVC. Requiring no domain expertise, additional visual modality is easy to collect and proves to be a strong signal for FGVC. Moreover, freely accessible global scale information such as temperature and digital elevation model (DEM) can easily be incorporated into our framework. We hope that this study paves the way for innovative future methods of species distribution modeling using deep learning.

References

[1] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

[2] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

[3] G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona, and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 595–604, 2015.

[4] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[5] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[6] S. Reed, Z. Akata, H. Lee, and B. Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 49–58, 2016.

[7] A. Dubey, O. Gupta, P. Guo, R. Raskar, R. Farrell, and N. Naik, "Pairwise confusion for fine-grained visual classification," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 70–86, 2018.

[8] Z. Yang, T. Luo, D. Wang, Z. Hu, J. Gao, and L. Wang, "Learning to navigate for fine-grained classification," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 420–435, 2018.

[9] Q. Diao, Y. Jiang, B. Wen, J. Sun, and Z. Yuan, "Metaformer: A unified meta framework for fine-grained recognition," *arXiv preprint arXiv:2203.02751*, 2022.

[10] Y. Zhang, H. Tang, and K. Jia, "Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data," in *Proceedings of the european conference on computer vision (ECCV)*, pp. 233–248, 2018.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

[12] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon, "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[13] S. Khanal, S. Sastry, A. Dhakal, and N. Jacobs, "Learning tri-modal embeddings for zero-shot soundscape mapping," *British Machine Vision Conference*, 2023.

[14] A. Dhakal, A. Ahmad, S. Khanal, S. Sastry, and N. Jacobs, "Sat2cap: Mapping fine-grained textual descriptions from satellite images," *arXiv preprint arXiv:2307.15904*, 2023.

[15] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pp. 348–367, Springer, 2022.

[16] Z. Tang, J. Cho, Y. Nie, and M. Bansal, "Tvlt: Textless vision-language transformer," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9617–9632, 2022.

[17] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[19] D. Jacob, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, p. 2, 2019.

[20] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, 2022.

[21] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multi-scale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

[22] X. Geng, H. Liu, L. Lee, D. Schuurams, S. Levine, and P. Abbeel, "Multimodal masked autoencoders learn transferable representations," *arXiv preprint arXiv:2205.14204*, 2022.

[23] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," *arXiv preprint arXiv:2207.13532*, 2022.

[24] C.-Z. Lu, X. Jin, Z. Huang, Q. Hou, M.-M. Cheng, and J. Feng, "Cmae-v: Contrastive masked autoencoders for video action recognition," *arXiv preprint arXiv:2301.06018*, 2023.

[25] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "Cmid: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[26] Z. Li, Y. Yang, X. Liu, F. Zhou, S. Wen, and W. Xu, "Dynamic computational time for visual attention," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 1199–1209, 2017.

[27] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4438–4446, 2017.

[28] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 5209–5217, 2017.

[29] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5012–5021, 2019.

[30] P.-Y. Chou, Y.-Y. Kao, and C.-H. Lin, "Fine-grained visual classification with high-temperature refinement and background suppression," *arXiv preprint arXiv:2303.06442*, 2023.

[31] X. Yu, Y. Zhao, and Y. Gao, "Spare: Self-supervised part erasing for ultra-fine-grained visual categorization," *Pattern Recognition*, vol. 128, p. 108691, 2022.

[32] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[33] F. A. Breiki, M. Ridzuan, and R. Grandhe, "Self-supervised learning for fine-grained image classification," *arXiv preprint arXiv:2107.13973*, 2021.

[34] X. Yang, J. Hu, Z. Wang, F. Xu, and L. Zhu, "Self-supervised fine-grained image classification via progressive global disturbance," in *Proceedings of the 4th International Conference on Computer Science and Software Engineering*, pp. 119–125, 2021.

[35] Y. Shu, A. van den Hengel, and L. Liu, "Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11392–11401, 2023.

[36] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, and L. Bourdev, "Improving image classification with location context," in *Proceedings of the IEEE international conference on computer vision*, pp. 1008–1016, 2015.

[37] G. Mai, N. Lao, Y. He, J. Song, and S. Ermon, "Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations," *International Conference on Machine Learning*, 2023.

[38] G. Chu, B. Potetz, W. Wang, A. Howard, Y. Song, F. Brucher, T. Leung, and H. Adam, "Geo-aware networks for fine-grained recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[39] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10181–10190, 2021.

[40] C. Greenwell, S. Workman, and N. Jacobs, "What goes where: Predicting object distributions from above," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4375–4378, IEEE, 2018.

[41] T. Salem, S. Workman, and N. Jacobs, "Learning a dynamic map of visual appearance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12435–12444, 2020.

[42] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10181–10190, 2021.

[43] Y. Xu, Q. Zhang, J. Zhang, and D. Tao, "Vitae: Vision transformer advanced by exploring intrinsic inductive bias," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28522–28535, 2021.

[44] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer towards remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[45] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2019.

[46] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *International Conference on Learning Representations*, 2017.

[47] S. G. Müller and F. Hutter, "Trivialaugment: Tuning-free yet state-of-the-art data augmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 774–782, 2021.

[48] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.

[49] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[50] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[52] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.