

A Self-Supervised Approach to Land Cover Segmentation

Charles Moore^{1,*} and Dakota Hester^{2,3,*}

¹Department of Computer Science and Engineering, Mississippi State University, Starkville, MS 39759, USA

²Department of Sustainable Bioproducts, Mississippi State University, Starkville, MS 39759, USA

³Department of Agricultural and Biological Engineering, Mississippi State University, Starkville, MS 39759, USA

*Corresponding authors: {cam1271,dh2306}@msstate.edu

Abstract

Land use/land cover change (LULC) maps are integral resources in earth science and agricultural research. Due to the nature of such maps, the creation of LULC maps is often constrained by the time and human resources necessary to accurately annotate satellite imagery and remote sensing data. While computer vision models that perform semantic segmentation to create detailed labels from such data are not uncommon, little research has been done on self-supervised and unsupervised approaches to labelling LULC maps without the use of ground-truth masks. Here, we demonstrate a self-supervised method of land cover segmentation that has no need for high-quality ground truth labels. The proposed deep learning employs a frozen pre-trained ViT backbone transferred from DINO in a STEGO architecture and is fine-tuned using a custom dataset consisting of very high resolution (VHR) satellite imagery. After only 10 epochs of fine-tuning, an accuracy of roughly 52% was observed across 5 samples, signifying the feasibility of self-supervised models for the automated labelling of VHR LULC maps.

1 Introduction

High-quality land cover use and land cover change (LU/LC) maps are critical to understanding and evaluating human interaction with natural resources and the effects of such interactions on both the natural environment and humans. LU/LC maps are critical in documenting climate change, de-

forestation, wildland fires, and atmospheric research [1, 2, 3]. Numerous advancements in remote sensing equipment, computational resources, and artificial have somewhat alleviated the need for large scale human-annotation of satellite imagery and remote sensing data, yet little research on scaling such techniques to very high resolution (VHR) data is still in its infancy [4, 5].

The goal of this work is to address the potential of self-supervised methods of training to overcome the labeled data bottleneck and to determine if the performance of such a model is applicable to the creation of deep learning techniques for the segmentation of land cover maps. By using a STEGO framework to distill correspondence similarities from a pre-trained model, our aim is to utilize existing models to determine the semantic information while evaluating feature correspondences in VHR satellite imagery to create detailed LULC maps.

2 Related Work

2.1 Land Use/Land Cover Change Maps

Land use/land cover change (LU/LC) maps detail natural resource usage

by a variety of researchers in various fields. LU/LC allow researchers to monitor developments in environmental interactions as well as model future changes in land use change [2, 6, 7]. However, development of high-quality large-scale land cover maps is labor-intensive and computationally expensive, and as such resolution is typically constrained: USGS NLCD 2019 maps land usage at a resolution of 30m. Due to rapid developments in remote sensing, the availability of very high resolution imagery and satellite data is now abundant, yet research on scaling algorithms used for labeling coarser data is still in its relative infancy [8]. With the emergence of recent self-supervised algorithms that possess the ability to learn semantically rich features without the need for high-quality annotations, further research is needed to evaluate their potential for labeling LULC maps at high resolution [9, 10].

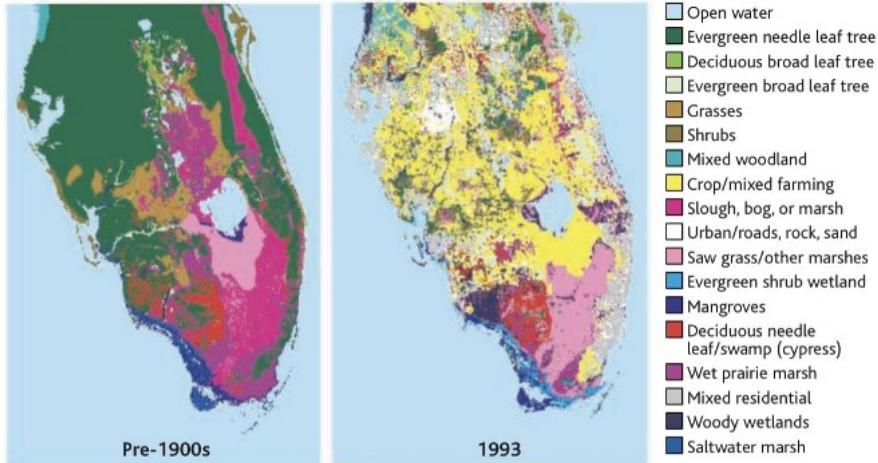


Figure 1: Example of large-scale LULC map of central and southern Florida, USA. LULC maps such as these are critical in evaluating and modeling human utilization of land and resources [1].

2.2 Semantic Segmentation

Hamilton *et al.* define semantic segmentation as “Semantic segmentation is the process of classifying each individual pixel of an image into a known ontology” [11]. Semantic segmentation is useful when all objects or features in an image and their location(s) within the image is necessary information. To date, most deep-learning semantic segmentation algorithms employ the use of convolutional neural networks (CNNs) in some fashion; CNNs are a special class of deep learning models powered predominantly by stacked convolutional layers that enable neural networks the ability to better capture low-level, mid-level, and high-level spatial features [12, 10]. In remote sensing, research on the use deep learning architectures to label land use and land change has emerged, but most implementations rely on access to hand-labeled annotations specific to the re-

gion of investigation of the desired resolution, which is not always feasible [13, 14].

2.3 Convolutional Neural Networks

CNNs primarily consist of convolutional layers, pooling layers, and fully connected layers arranged in such a hierarchy to allow a representation of low-level features (edges, lines, corners) to be used to learn more complex high-level features (faces, objects, etc.) [12]. CNNs are widely popular in most computer vision applications, with tasks ranging from image classification to optical flow, and as such are commonly used in the processing of spatial remote sensing data [15]. For semantic segmentation specifically, DeepLabv3+ (see figure 2) is one of the most popular frameworks for vision tasks that use supervised learning [16, 17].

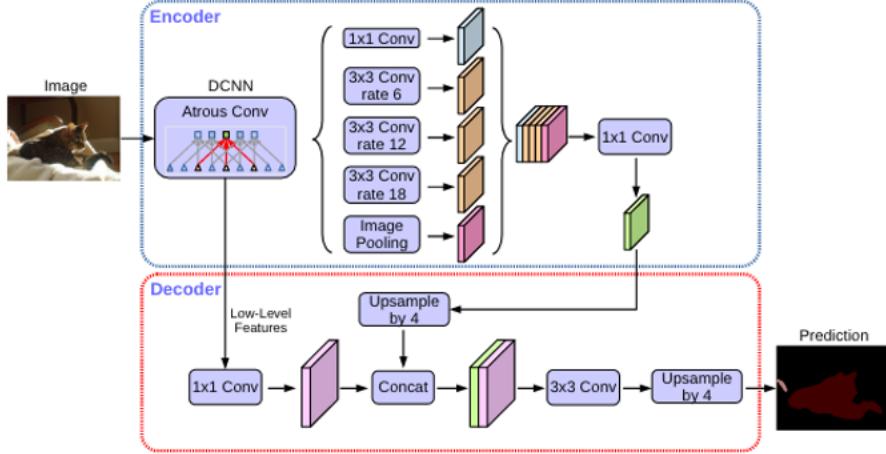


Figure 2: Simplified graph of DeepLabv3+ architecture. Note the extensive use of convolutional layers arranged in a hierarchy. Supervised architectures such as these perform well in scenarios where masked training data is of high-quality and abundant [17, 16].

2.4 Contrastive Learning

Contrastive learning is a framework for self-supervised learning that operates on the principle that representations of samples from similar classes should be similar, whereas representations of samples from disparate classes should be dissimilar. Many common contrastive learning algorithms enforce this principle with the use of data augmentation: a pair of subsamples is created by taking a single image as input (referred to as an "anchor") before randomly cropping, stretching, or resizing the image in such a way to create two dissimilar images that belong to the same class - otherwise known as a positive pair. Negative pairs consisting of two augmented images from different classes are also used in training with one exception; their representations should be dissimilar. A contrastive loss function attempts to max-

imize agreement between positive pairs and minimize agreement between negative pairs. Specifically, when classes are unknown, other augmented images in a given batch that are **not** derived from the original "anchor" image are treated as negative pairs. [18, 9].

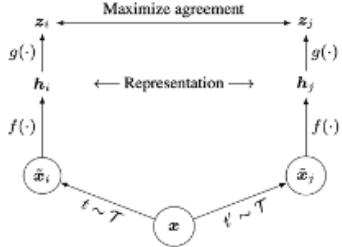


Figure 3: Graphical depiction of SimCLR - a popular contrastive learning framework. An input X is augmented into a positive pair. Loss is computed over the representations received from a multilayer perceptron (MLP) such to minimize the differences between the two representations during training [9].

2.5 Transformers in Computer Vision

Self-attention is an important component of state-of-the art models built for natural language processing (NLP). Implementing self-attention into computer vision models has the potential to allow deep learning models to learn features in an image in a global fashion rather than local representations confined to the receptive fields of convolutional layers. The most popular vision-based adaptations of these methods draw from Transformer-based models commonly applied in NLP domains such as BERT and GPT-3 with the goal of overcoming such limitations on learning long-range interactions [19, 20, 10, 21, 18].

Vision Transformer (ViT) is a family of self-attention-based architecture that divides an image into a series of patches that are then placed into a multi-head self attention block. Patches are flattened into a linear projection before input into a large model consisting of several stacked Transformer encoders - each consisting of global self-attention layers. Self-attention layers are then fed into multi-layer perceptrons (MLP) at the final stage of an encoding layer. Residual connections concatenate the input of the encoder to the output of the multi-head self attention block, and the output of said self attention blocks to the final output of the layer. Positional encodings are also inputted alongside each linearly embedded patch, but these encodings do not convey information regarding the position of each embedded patch before training - meaning spatial information and relationships must be learned from scratch (as opposed to CNNs). Regardless, the performance of ViTs is competitive with or exceeds that of state-of-the-art CNN methods for classification when trained on several benchmark datasets while being considerably more computationally efficient in terms of resource use during pre-training. However Big Transformer (BiT) - a family of similar architecture that utilizes large CNNs during pre-training - yielded comparatively better results when the number of pre-training samples was reduced [21].

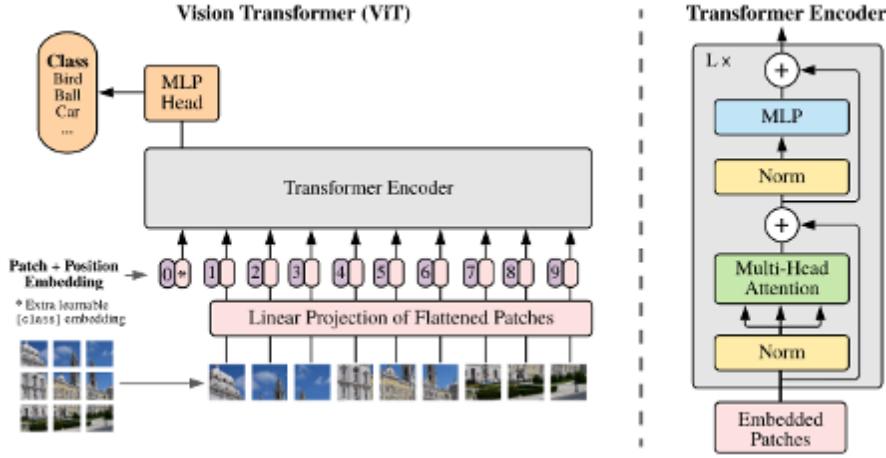


Figure 4: Graphical overview of basic ViT architecture. Note the multi-head self attention layer included in each transformer encoder layer [21].

2.6 DINO

Whilst ViTs have shown to be of similar or higher performance compared to state-of-the-art CNNs, their large size and need for huge amounts of labeled pre-training data make them relatively impractical for implementations in situations where computational resources and data availability are constraints. Caron *et al.* showed that a self-supervised approach to pre-training using contrastive learning showed considerably better performance on downstream segmentation tasks compared to models pre-trained in a fully supervised procedure. This approach, named **DINO** (**D**istilling knowledge with **no** labels) was able to learn features that were richer and more semantically meaningful compared to those learned by ViTs trained in a fully supervised fashion - making it suitable for usage in various vision tasks beyond classification. DINO’s ability to learn rich features from samples without an

notation stems from its use of knowledge distillation to train a new model against an ensemble of previous models using augmented pairs from the same sample [22, 10].

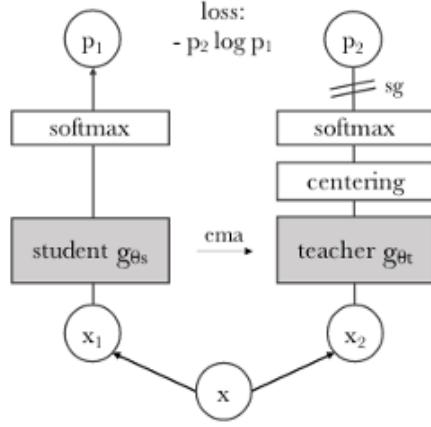


Figure 5: Graphical overview of DINO training. Note how the training step utilizes a contrastive learning approach. Knowledge distillation is used to train a teacher and student model, here student model g_{θ_s} (ViT backbone at current epoch) is trained to match the output of teacher model g_{θ_t} (ensemble of ViT backbones at all previous epoch). Using such a knowledge distillation during training was shown to yield consistent high-quality targets from the teacher model [10].

2.7 STEGO

One key observation from DINO is that once features have been captured

in a self-supervised fashion during pre-training, correlations between learned features are consistent with semantic information found not only within the same image but across other images as well [10]. STEGO is an architecture recently introduced to perform unsupervised semantic segmentation by distilling feature correspondences (retrieved from a frozen pre-trained DINO ViT backbone) across and between samples into a lower dimensional representation using a simple MLP. It requires no fine-tuning due to DINO’s ability to capture rich features in pre-training - the outputs of the frozen DINO backbone are fed to a single segmentation head consisting of a simple MLP. By utilizing a basic but creative contrastive learning implementation where a contrastive loss function takes into account the feature correspondences in a pair of samples, this method has yielded breakthrough results on unsupervised semantic segmentation tasks. Further, the architecture is highly adaptable to various scenarios and domains as the architecture itself does not define a strict method for pre-training, thus any future advancements in vision models or pre-training paradigms can be easily patched in [11].

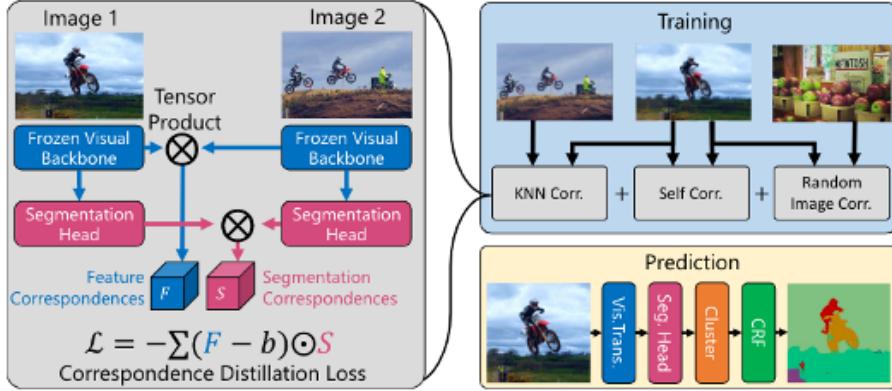


Figure 6: Graphical overview of STEGO model. Note the [11].

3 Materials & Methodology

3.1 Datasets

For this research, two datasets consisting of VHR satellite imagery and their corresponding masks were used in various capacities in training and evaluation. Whilst pre-training was done using DINO with data from ImageNet, parameters in the backbone ViT were transferred to the downstream STEGO model for fine-tuning. Ideally, a sufficiently large dataset of VHR imagery would be used to pre-train DINO whereas a smaller dataset would then be used to fine-tune. To the authors’ knowledge, no such publicly accessible dataset exists at this point in time.

3.1.1 DeepGlobe

The DeepGlobe dataset is a high-resolution satellite image dataset that consists of three challenge tracks: road extraction, Building Detection, and Land Cover Classification. We utilized the Land Cover Classification track,

which is in accordance with the aim of this project. Land cover classification is important for tasks such as: sustainable development, autonomous agriculture, and urban planning. The Land Cover Classification task is split into seven classes: urban, agriculture, rangeland, forest, water, barren, and unknown. This is defined as a multi-class segmentation task.

The DeepGlobe Land Cover Classification consists of 1,146 images at a spatial resolution of 50 centimeters per pixel. Inputs were resized to 256x256x3 with a resultant spatial resolution of 5 meters per pixel [23].

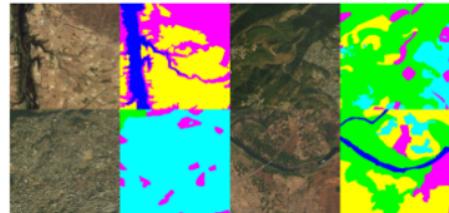


Figure 7: Selected images and masks from DeepGlobe Land Cover Challenge Dataset [23].

3.1.2 LandCover.ai

Similar to the DeepGlobe dataset (Land Cover Classification task), the LandCover.ai dataset is an RGB manually-annotated image dataset based on satellite images for land cover classification. The images from this dataset come from Central Europe with most images coming from Poland. The landscape of Poland is primarily dominated by mixed forests and agrarian areas. This dataset is divided into four different classes: buildings, woodlands, water, and road for land cover generality and usefulness for public administrations. However, we discarded the need for labels from this dataset as we only used the source images to train the STEGO model.

In the Land Cover Classification track, the labels are only publicly available for the train splits. We decided to combine the validation and test source images, along with images from the LandCover.ai dataset, to create a pseudo-train dataset. This allows us to obtain accuracy values from the DeepGlobe train set.

The LandCover dataset consists of 10,604 images with spatial resolutions ranging from 50 centimeters per pixel to 25 centimeters per pixel. After resizing to 256x256x3, the resultant spatial resolution ranged from 50-100 centimeters per pixel [24].

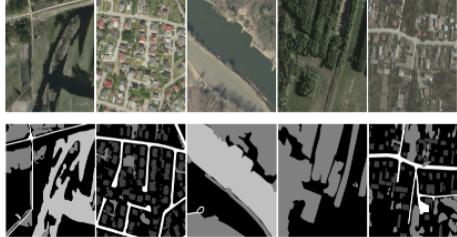


Figure 8: Selected images and masks from LandCover.ai dataset [24].

3.2 Model Configuration

In order to achieve best results, we opted to utilize various vision transformers that were pre-trained by the DINO paradigm. It is important to note that the vision transformers were pre-trained on the ImageNet dataset. This dataset includes 14+ million images. This allows the vision transformer to see a diverse array of pre-training data and allows the potential for higher segmentation accuracy compared to models pre-trained on smaller datasets [10]. In order to train the STEGO segmentation decoder, we utilized a set batch size of 16, learning rate of 0.0001, an Adam optimizer, and momentum constraints preset by the STEGO authors. It is also important to note that the learning rate is scalable, as the STEGO architecture allows for such a mechanism [11]. We set training and inference images to a set size of 256 x 256 to allow for the model to train efficiently while relaxing memory constraints and also prevent the loss of valuable global information. We limit the model to train on 200 images with around 10 epochs to train due to compute constraints.

4 Results

Given the limited number of epochs and small training dataset, the results are greater than expected. However, we expect the quality of the features from the vision transformer and the possibility of the model reaching an optimum quickly to be the reasoning behind this. In most, if not all, of the inference images, STEGO performs well in identifying segmentation areas, but does seem to over-represent some classes, resulting in a class imbalance. We were able to observe the model achieving a 52% accuracy across 5 test samples with the given model hyper-

parameters (see figure 9). We believe that with training data and better optimized hyperparameters, a modified STEGO model can achieve state-of-the-art performance while simply utilizing a pre-trained model from DINO. It is important to note that this result was achieved without a domain focused pre-trained dataset. We can expect the model to perform better using the weights from a DINO backbone pre-trained with a sufficiently large dataset of consistent high-resolution satellite imagery - such a dataset is not publicly available to the authors' knowledge.

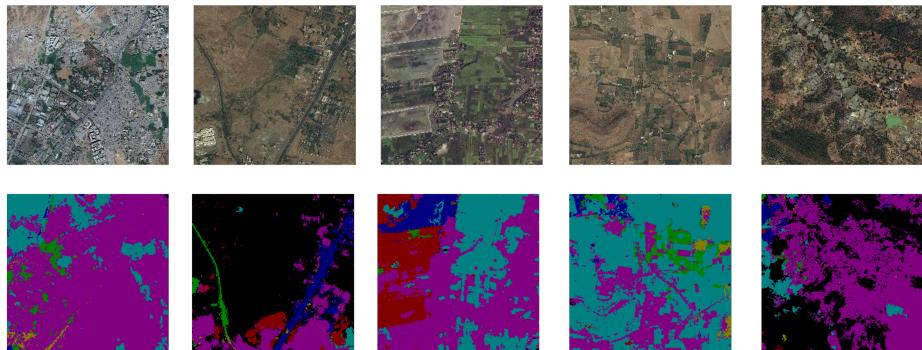


Figure 9: Images segmented using proposed model. Images on the top were segmented using the STEGO model after fine-tuning, masks on the bottom were the resultant masks from the segmentation component of the model. Despite only 10 epochs of fine-tuning and no ground truth labels to rely on, the model was still able to discern differences in spatial features to yield a fairly accurate map of land cover.

5 Discussion

We found that utilizing a pre-trained DINO vision transformer as a backbone for the STEGO decoder can yield decent results in semantic segmentation. We were able to obtain 52% accuracy with a small batch size and

minimal training images. Considering that accuracy results over 70% are typically considered acceptable for models that utilize fully supervised training in the remote sensing research domain [25, 26, 27], these results are promising. We believe that scaling up the datasets used in both training phases

will not only yield better results but will also improve the model’s robustness. Although STEGO is proven to be a good option for semantic segmentation, there exist potential improvements to the architecture and methodology. One such potential modification would be to incorporate the MASK DINO framework into the STEGO architecture to take advantage of potentially richer learned semantic features [28]. Another potential investigation would be the performance of the MoBY architecture with the improved SwinV2 Vision Transformer variant on VHR satellite imagery segmentation tasks due to SwinV2’s ability to handle high-resolution samples [29, 30]. In addition to LULC segmentation, there are various similar domains where such work can be applied to. Large scale terrain segmentation is an important area of study in relation to land cover classification. An accurate model for terrain segmentation can easily assist tasks in robotics such as exploration. This provides a reason to scale up from aerial images to land masses and remains a widely researched topic in computer vision. Another area of investigation that stems from this project is the concept of using real world data and observations to fix model inaccuracies during a robotics task such as exploration or path planning. Successfully incor-

porating real world data could have the effect of needing less training data to build models, thus saving time and resources training models.

6 Conclusion

Land cover classification is still an important problem to date, and is an important task in the field of computer science for agricultural and commercial reasons [23, 31]. In this study, we have shown that the combination of DINO and STEGO has the potential to be a feasible alternative to supervised learning models. Contrastive learning is relatively new in the area of computer vision but has the ability to generate robust models that can be applied to various computer vision tasks: such models can provide many benefits, and we believe that smaller models trained with contrastive learning can achieve performance competitive with supervised learning methods. Self-supervised learning offers a good alternative for supervised training on unreliable training datasets [29, 18], and we have shown that self supervision can do an acceptable job at distinguishing different land cover classes in VHR aerial images. As such, further research could potentially reveal solutions to problems in earth science, climatology, and agriculture.

7 References

References

- [1] R. A. Pielke, “Land Use and Climate Change,” *Science*, vol. 310, no. 5754, pp. 1625–1626, Dec. 2005.

- [2] J. J. Feddema, K. W. Oleson, G. B. Bonan, L. O. Mearns, L. E. Buja, G. A. Meehl, and W. M. Washington, “The Importance of Land-Cover Change in Simulating Future Climates,” *Science*, vol. 310, no. 5754, pp. 1674–1678, Dec. 2005.
- [3] D. P. Roy, H. Huang, R. Houborg, and V. S. Martins, “A global analysis of the temporal availability of PlanetScope high spatial resolution multispectral imagery,” *Remote Sensing of Environment*, vol. 264, p. 112586, Oct. 2021.
- [4] M. A. Wulder, N. C. Coops, D. P. Roy, J. C. White, and T. Hermosilla, “Land cover 2.0,” *International Journal of Remote Sensing*, vol. 39, no. 12, pp. 4254–4284, Jun. 2018.
- [5] S. N. MohanRajan, A. Loganathan, and P. Manoharan, “Survey on Land Use/Land Cover (LU/LC) change analysis in remote sensing and GIS environment: Techniques and Challenges,” *Environmental Science and Pollution Research*, vol. 27, no. 24, pp. 29 900–29 926, Aug. 2020.
- [6] G. Zeleke and H. Hurni, “Implications of Land Use and Land Cover Dynamics for Mountain Resource Degradation in the Northwestern Ethiopian Highlands,” *Mountain Research and Development*, vol. 21, no. 2, pp. 184–191, May 2001.
- [7] R. A. Pielke, A. Pitman, D. Niyogi, R. Mahmood, C. McAlpine, F. Hossain, K. K. Goldewijk, U. Nair, R. Betts, S. Fall, M. Reichstein, P. Kabat, and N. de Noblet, “Land use/land cover changes and climate: Modeling analysis and observational evidence,” *WIREs Climate Change*, vol. 2, no. 6, pp. 828–850, Nov. 2011.
- [8] C. Homer, J. Dewitz, S. Jin, G. Xian, C. Costello, P. Danielson, L. Gass, M. Funk, J. Wickham, S. Stehman, R. Auch, and K. Riitters, “Conterminous United States land cover change patterns 2001–2016 from the 2016 National Land Cover Database,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 184–199, Apr. 2020.
- [9] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” Jun. 2020.
- [10] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, “Emerging Properties in Self-Supervised Vision Transformers,” May 2021.
- [11] M. Hamilton, Z. Zhang, B. Hariharan, N. Snavely, and W. T. Freeman, “Unsupervised semantic segmentation by distilling feature correspondences,” 2022.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 2015.

- [13] K.-A. Nguyen and Y.-A. Liou, “Mapping global eco-environment vulnerability due to human and nature disturbances,” *MethodsX*, vol. 6, pp. 862–875, 2019.
- [14] V. Martins, D. Roy, H. Huang, L. Boschetti, H. Zhang, and L. Yan, “Deep learning high resolution burned area mapping by transfer learning from Landsat-8 to PlanetScope,” *Remote Sensing of Environment*, vol. 280, p. 113203, Oct. 2022.
- [15] V. S. Martins, A. L. Kaleita, B. K. Gelder, H. L. da Silveira, and C. A. Abe, “Exploring multiscale object-based convolutional neural network (multi-OCNN) for remote sensing image classification at high spatial resolution,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 56–73, Oct. 2020.
- [16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation,” Dec. 2017.
- [17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation,” Aug. 2018.
- [18] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A Survey on Contrastive Self-supervised Learning,” Feb. 2021.
- [19] Z. Fu, “Vision Transformer: Vit and its Derivatives,” May 2022.
- [20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” Aug. 2021.
- [21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Jun. 2021.
- [22] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015.
- [23] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, “DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 172–17209.
- [24] A. Boguszewski, D. Batorski, N. Ziemb-Jankowska, T. Dziedzic, and A. Zambrzycka, “LandCover.ai: Dataset for Automatic Mapping of Buildings, Woodlands, Water and Roads from Aerial Imagery,” Apr. 2022.

- [25] B. Chen, M. Xia, and J. Huang, “MFANet: A Multi-Level Feature Aggregation Network for Semantic Segmentation of Land Cover,” *Remote Sensing*, vol. 13, no. 4, p. 731, Feb. 2021.
- [26] K. Karra, C. Kontgis, Z. Statman-Weil, J. C. Mazzariello, M. Mathis, and S. P. Brumby, “Global land use / land cover with Sentinel 2 and deep learning,” in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. Brussels, Belgium: IEEE, Jul. 2021, pp. 4704–4707.
- [27] M. B. A. Gibril, M. O. Idrees, K. Yao, and H. Z. M. Shafri, “Integrative image segmentation optimization and machine learning approach for high quality land-use and land-cover mapping using multisource remote sensing data,” *Journal of Applied Remote Sensing*, vol. 12, no. 01, p. 1, Mar. 2018.
- [28] F. Li, H. Zhang, H. xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, “Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation,” Nov. 2022.
- [29] Z. Xie, Y. Lin, Z. Yao, Z. Zhang, Q. Dai, Y. Cao, and H. Hu, “Self-Supervised Learning with Swin Transformers,” May 2021.
- [30] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, “Swin Transformer V2: Scaling Up Capacity and Resolution,” Apr. 2022.
- [31] C. Giri, B. Pengra, J. Long, and T. Loveland, “Next generation of global land cover characterization, mapping, and monitoring,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 25, pp. 30–37, Dec. 2013.