

Leveraging Masked Autoencoders For Few-Shot Transformer-Based Semantic Segmentation

Akansha Verma, Ayush Supakar, Dakshi Vashishtha, Riya Garg and Ashwani Kumar,
Department of Computer Science, Bennett University, Greater Noida, India

Abstract—The semantic segmentation of satellite imagery has broad applications in urban planning, disaster relief, and agriculture. However, few-shot learning strategies are highly desirable since it is expensive to create large-scale pixel-level annotations. In this work, we propose a framework which uses a ViT-based U-Net for downstream segmentation and Masked Autoencoders (MAE) for pre-training. The MAE encoder learns robust representations by self-supervised reconstruction of randomly masked patches without the need for labeled pre-training data. For semantic segmentation, these pre-trained features are further refined with a small number of annotated samples. In the end, our model achieves a mean IoU of 0.1911 and a Dice coefficient of 0.8963, confirming that self-supervised learning methods, such as MAE, have great potential for alleviating this heavy dependence on large labeled datasets.

Index Terms—Self-supervised learning, Vision Transformer, Satellite image segmentation, Few-shot learning, Masked Autoencoder, Remote sensing, Building detection

1. INTRODUCTION

The goal of semantic segmentation, a basic computer vision task, is to give each pixel in an image a semantic label. Segmentation makes it possible to precisely classify land cover, analyze urban infrastructure, monitor agriculture, and manage disasters in the context of remote sensing. Although high-resolution satellite imagery offers important insights into socio-economic and environmental processes, it is still difficult to extract structured information from this type of imagery. Conventional deep learning approaches, particularly convolutional neural networks (CNNs), have achieved significant success in semantic segmentation. Architectures such as U-Net and DeepLab have set strong benchmarks across several domains. However, these models typically require large amounts of annotated data at the pixel level for effective training. In remote sensing, generating such annotations can be costly, time consuming, and often unfeasible, particularly when considering diverse geographical regions or specific land cover categories. This data scarcity motivates the exploration of few-shot learning techniques, where models are expected to generalize well even with limited labeled data. Recently, *Vision Transformers (ViTs)* have emerged as powerful alternatives to CNNs by modeling long-range dependencies through self-attention mechanisms Fig.1. However, training ViTs from scratch on limited datasets is inefficient and prone to overfitting. To overcome this limitation, self-supervised pretraining strategies have gained attention. In particular, *Masked Autoencoders (MAEs)* have demonstrated remarkable ability in learning robust image representations without labeled

supervision. By reconstructing missing image patches, MAEs force the encoder to capture contextual semantics, which can then be transferred to downstream tasks. In this work, we propose a novel framework that leverages MAE-pretrained Vision Transformers for few-shot semantic segmentation in remote sensing. In particular, we combine a lightweight U-Net-style decoder for segmentation with a ViT encoder that was pre-trained using MAE. While the decoder reconstructs fine-grained segmentation masks, the pretrained encoder offers rich feature representations. Through extensive experiments on remote sensing datasets, we demonstrate that our method achieves higher mean Intersection-over-Union (mIoU), Dice coefficient, and pixel accuracy compared to traditional baselines, thereby significantly improving segmentation performance under data-scarce conditions.

The main contributions of this paper are as follows:

- We introduce a MAE-ViT-U-Net hybrid framework for few-shot semantic segmentation in remote sensing.
- We show that MAE pretraining enhances the feature quality of ViTs, enabling effective transfer to segmentation with minimal labeled data.
- We provide comprehensive quantitative and qualitative evaluations, demonstrating improvements in IoU, Dice, and pixel accuracy over standard CNN-based baselines.
- We highlight the practical relevance of our approach in real-world scenarios where annotated data is scarce, making it suitable for scalable deployment in remote sensing applications.

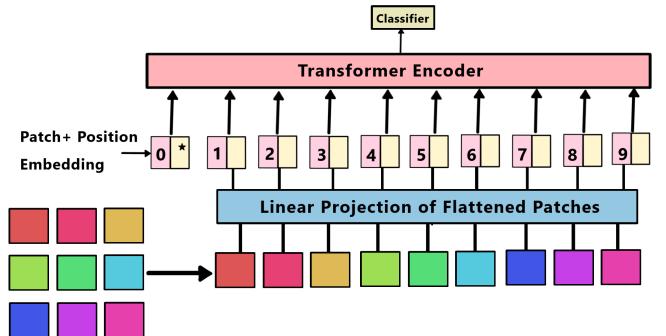


Fig. 1: Simple architecture for ViT.

2. LITERATURE REVIEW

A. Satellite image segmentation

From traditional machine learning techniques to advanced deep learning strategies, semantic segmentation of satellite imagery has been a popular area of research. Early techniques relied on spectral analysis and texture-based features with conventional classifiers like Random Forests and Support Vector Machines (SVM), as demonstrated by Saha *et al.* [1] offered interpretability but often struggled to capture multi-scale features and spatial patterns in high-resolution satellite images.

1) *Traditional Deep Learning Approaches:* : Convolutional Neural Networks (CNNs) started it for satellite image analysis with Fully Convolutional Networks (FCNs) establishing end-to-end learning for dense prediction which then formed the basis for everything. Ronneberger et al.'s U-Net architecture [2] introduced skip connections to preserve the fine-grained spatial information. Chen et al.'s DeepLab series [3] then employed atrous convolutions and pyramid pooling to perceive features at varying scales. It is a big deal while detecting objects of all shapes and sizes within high-resolution satellite imagery. To address challenges in maritime remote sensing, Li et al. proposed DeepUNet [4], a refined U-Net architecture for pixel-wise sea-land segmentation, which showed enhanced performance over the standard U-Net. Recently, attention mechanisms have been added to more accurately capture how different pieces of a huge satellite image are related to one another, something which Zhang et al. have investigated [5].

2) *Transfer Learning in Remote Sensing:* Tagged satellite data is rare and costly. Therefore, transfer learning is routine practice, typically starting with a model trained on a massive natural image dataset such as ImageNet and then fine-tune it. Penatti et al. [6] discovered that pretraining with ImageNet provides a reasonable start, but the domain difference between objects of daily life and Earth observation is significant. Some research challenges this dependency on natural images. Neumann et al. [7], showed that pretraining on large, unlabeled satellite dataset actually performs better than ImageNet. It is just a testament to how much value domain-specific pre training holds and why self-supervised approaches are gaining such significance.

3) *Multi-spectral and Multi-temporal Analysis:* Satellite data is nuanced, containing several spectral bands. Scientists have exploit this, such as Audebert et al. [8], who merged RGB and near infrared data for land cover mapping. Kemker et al. [9] also explored how various blends modify a model's performance. Aside, there is also the factor of time. Observing the series of images can uncover dynamic changes, such as crop growth. Rußwurm et al. [10] applied LSTMs on Sentinel-2 time series data to accurately capture seasonal changes. However, this method requires large, well-labeled datasets, which are difficult to obtain.

B. Self-Supervised Learning

1) *Contrastive Learning Methods:* This method learns by comparing positive and negative examples pairs. SimCLR

by Chen et al. [11] showed that contrastive learning can match supervised pretraining on ImageNet. MoCo by He et al. [12] introduced momentum-based learning, which uses large negative sets. BYOL by Grill et al. [13] proved that contrastive learning could succeed without any negative samples. These methods were successful in natural image contexts but have been less explored.

2) *Masked Image Modeling:* : Inspired by NLP, MAE by He et al. [14] masked random patches of input images and train models to reconstruct the missing content. BEiT by Bao et al. [15] combined masked image modeling with discrete visual tokens. More recently, ConvMAE by Gao et al. [16] adjusted masked image modeling for convolutional architectures.

3) *Self-Supervised Learning in Remote Sensing:* The use of SSL is rapidly growing. SeCo by Manas et al. [17] introduced seasonal contrast using the temporal consistency of locations across seasons. SSL4EO by Wang et al. [18] benchmarked self supervised methods on Earth observation data. However, many existing SSL methods for satellite imagery focus on image level tasks like classification rather than dense prediction tasks such as segmentation. Our research addresses this gap by specifically creating self-supervised pretraining for segmentation applications.

C. Vision Transformers

The introduction of Vision Transformer (ViT) by Dosovitskiy et al. [19] marked a significant change in computer vision proving that transformer architectures can achieve top performance on image recognition tasks without relying on convolutional biases.

1) *Transformer Architectures for Vision:* Following ViT's release, numerous variants have been proposed to enhance efficiency and performance. Swin Transformer by Liu et al. [20] presented hierarchical transformer architectures with shifted windowing, achieving excellent results across various computer vision tasks. PVT by Wang et al. [21] and Twins by Chu et al. [22] explored pyramid structures within transformer architectures which enables multi-scale feature extraction similar to what CNNs offer.

2) *Transformers for Semantic Segmentation:* Several studies have adapted transformer architectures for semantic segmentation. SETR by Zheng et al. [23] was among the first to use pure transformers for segmentation, utilizing ViT as the backbone with various decoding strategies. SegFormer by Xie et al. [24] proposed a straightforward yet effective transformer based segmentation model that does not rely on positional encodings. OneFormer by Jain et al. [25] unified instance, semantic, and panoptic segmentation demonstrating flexibility of transformer architectures for dense prediction .

D. Few-Shot and Transfer Learning

Few-shot learning is important for applications with scarce labeled data. Meta-learning methods such as MAML by Finn et al. [26] aim to adapt to new tasks with just a few examples. Prototypical networks by Snell et al. [27] learn representations by calculating distances to class prototypes. FSS-1000 by

Li et al. [28] established benchmark datasets and baseline methods for few-shot segmentation. HSNet by Min et al. [29] introduced hypercorrelation squeeze networks to improve outcomes. Most approaches depend on traditional transfer learning or domain adaptation. Our research contributes to this area by showing how self-supervised pretraining can enable effective few-shot segmentation in satellite imagery.

3. PROPOSED METHODOLOGY

This approach leverages the powerful representation capabilities of Masked Autoencoders (MAE) and the spatial accuracy of a U-Net decoder based on a ViT encoder as shown in Fig.(2). In general, the framework is designed to reduce reliance on large labeled datasets while achieving high segmentation performances with few examples.

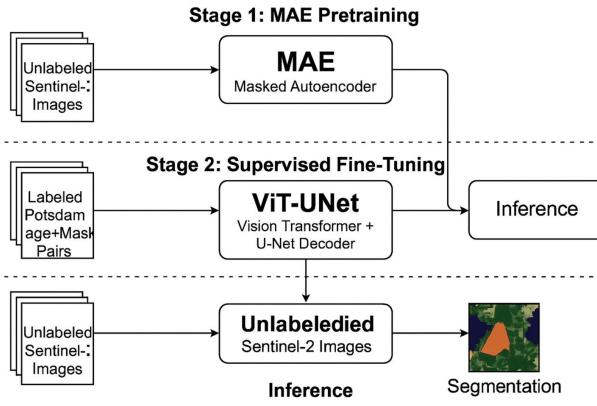


Fig. 2: Proposed methodology for MAE-pretrained Vision Transformer based semantic segmentation.

A. Overview

The key idea behind this architecture is that it decouples the process of representation learning from further fine-tuning based on tasks. It learns helpful visual representations in a self-supervised manner in the first stage from unlabeled Sentinel-2 (EuroSAT) satellite images [30], [31]. In the next stage, the pretrained representations are transferred to a U-Net-based segmentation model, where they get fine-tuned over the ISPRS Potsdam dataset [32] with scarce labeled data for 6-class multi-class segmentation. This design enables the two-stage architecture to let the encoder capture the general spatial, spectral, and structural patterns from a big set of unlabeled data and effectively adapt these representations for dense prediction tasks with very minimal supervision. The overall pipeline is given in Fig. 2.

B. Stage I – Self-Supervised Pretraining with Masked Autoencoder

1) *Purpose:* The self-supervised pretraining stage relies on the Masked Autoencoder approach [14], training a model to rebuild missing parts of an image after randomly concealing a large percentage of input patches. This process forces the

encoder to learn strong and context-aware representations that capture global dependencies, spectral relationships, and geographic structures found in satellite imagery.

We use a Vision Transformer (ViT) encoder for processing visible patches of Sentinel-2 images. An input image is divided into non-overlapping patches, and a random subset of them is masked. Only visible patches are fed into the encoder while the decoder is responsible for reconstructing the missing portions.

The learning objective minimizes the mean squared reconstruction loss between the reconstructed \tilde{I}_{mask} and ground truth I_{mask} masked patches:

$$L_{\text{MAE}} = \|I_{\text{mask}} - \tilde{I}_{\text{mask}}\|^2$$

Through this task, the encoder learns domain-specific representations without any manual annotations.

2) Training Configuration:

- **Dataset:** EuroSAT Sentinel-2 (RGB bands only)
- **Epochs:** 8
- **Batch Size:** 4
- **Optimizer:** Adam
- **Learning Rate:** 1×10^{-3}
- **Loss Function:** Mean Squared Error (MSE)
- **Gradient Clipping:** Value = 1.0
- **Precision:** Mixed Precision (AMP)
- **Augmentations:** Random flips, rotations, and brightness adjustments using Albumentations
- **Environment:** Google Colab GPU
- **Seed:** 42 for reproducibility

This stage leads to a pretrained ViT encoder that captures the spectral and spatial variability in satellite images. The model forms the base for the subsequent segmentation task.

C. Stage II: Supervised Fine-Tuning with ViT-U-Net Hybrid Fine-tuning step

The features learned during pretraining are fine-tuned to the labeled ISPRS Potsdam data. This dataset has pixel-wise labels for six classes: we treat the problem as binary segmentation, separating the buildings from the background into positive and negative classes.

This model fine-tunes on a hybrid loss function that incorporates Binary Cross-Entropy (BCE) and Dice Loss. This function balances pixel-wise classification with mask overlap optimization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}$$

where

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2|P \cap G|}{|P| + |G|}$$

Here, P denotes the predicted mask, and G represents the ground truth.

1) Optimization Strategy:

- **Epochs:** 10
- **Batch Size:** 2
- **Optimizer:** Adam
- **Learning Rate:** 1×10^{-4}
- **Loss Function:** BCE + Dice
- **Gradient Clipping:** 1.0
- **Early Stopping:** Patience = 10 epochs
- **Precision:** Mixed precision (PyTorch AMP)

We used gradient accumulation to handle memory limitations. Early stopping started when the validation loss did not improve over consecutive epochs.

The process of fine-tuning enhances this pretrained encoder by allowing it to understand the precise spatial boundaries and variations in texture necessary for accurately detecting buildings.

D. Evaluation Protocol

Quantitative and qualitative approaches were considered in assessing the performance of the segmentation model.

Quantitative Metrics:

- **IoU:** Intersection over Union
- **Dice Coefficient**
- **Pixel Perfect**

These metrics were computed on a separate validation set drawn from the Potsdam dataset. MAE-pretrained encoders, therefore, achieve higher IoU and Dice scores compared to baseline CNNs that are trained from scratch.

Qualitative Analysis: Predicted masks' visualizations were created with Matplotlib. They indicated better definition of boundaries and fewer false detections. Confusion matrices and per-class IoU plots have been used to check model consistency and performance by class.

4. RESULTS AND DISCUSSION

We evaluate the performance of the proposed pipeline for a two-stage training process: Masked Autoencoder pretraining followed by ViT-UNet fine-tuning. Two important datasets were employed in the whole process: the Sentinel-2 EuroSAT dataset for self-supervised pretraining and the ISPRS Potsdam dataset for supervised fine-tuning and evaluation.

For the pretraining stage, we used the EuroSAT dataset. This dataset provides around 27,000 unlabeled 64x64 pixel RGB images from 34 European countries. We used these images without their labels to train the MAE model. In the downstream segmentation task, we have used the ISPRS Potsdam dataset. The dataset includes very high-resolution (5cm) aerial imagery over Potsdam, Germany, and comprises 38 tiles in total. For our building segmentation task, we turned the six semantic classes of this dataset (e.g., Impervious Surface, Building, Tree) into a binary classification problem, defining "Building" as the positive class and all other pixels as the negative "Background" class. These results indicated that this pretraining stabilizes the process of feature learning, while fine-tuning improves the segmentation accuracy on labeled data.

A. Pretraining Performance

First, in this work, MAE is pre-trained for eight epochs using the Sentinel-2 images without labels. We observe a progressive decay of the loss from 0.5891 at epoch 0 to 0.2500 at epoch 6 as shown in Fig.(3), which demonstrates that MAE learns effective visual representation through reconstruction. The reduced loss indicates that the model has become much more capable of reconstructing masked areas in an image and, therefore, capturing the spatial relationship across patches. The smooth convergence further indicates this pretraining stage places the encoder of the Vision Transformer in a more robust setting for downstream segmentation tasks.

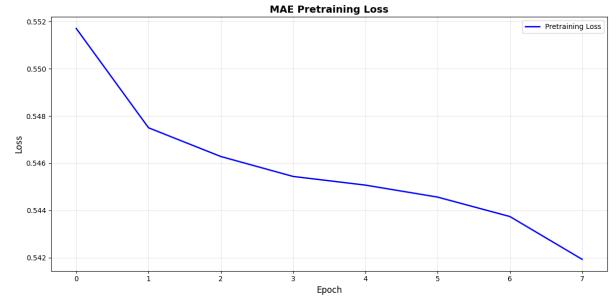


Fig. 3: Pretraining performance of the MAE-based ViT model over training epochs.

B. Fine-Tuning Performance

After transferring the pre-trained weights of the MAE encoder, fine-tuning was done on the labeled Potsdam image-mask pairs. The training and validation loss curves described in Fig. 4 outline a converging pattern: the training loss decreases from 1.23 to 0.73, while the validation loss goes from 1.19 to 0.63 up to epoch 9. This convergence pattern shows that the learning is quite stable with less overfitting due to added regularization through data augmentation and combined loss, Cross-Entropy + Dice.

Both IoU and Dice improve steadily with the epochs, ending with a final mIoU of 0.1911 and Dice of 0.8963 accordingly. Such a high Dice means that the model has captured the vast majority of the target regions correctly, while a moderate mIoU reflects that there are some overlap errors among classes, likely due to inter-class spectral similarity in the dataset.

C. Per-Class IoU Analysis

A class-wise IoU distribution, plotted in Fig. 5, shows that the IoU of Class 0 and Class 5 is 0.687 and 0.460, respectively, while others have an IoU close to zero. These results indicate heavy imbalance in the dataset, with the model managing effectively to learn only from dominant categories of land cover but failing for minority classes or those which were visually similar. This could also be because of dataset imbalance or perhaps due to the number of fine-tuning epochs being limited. Class-balancing strategies or focal loss may be further used for experimentation that could help in improving the results.

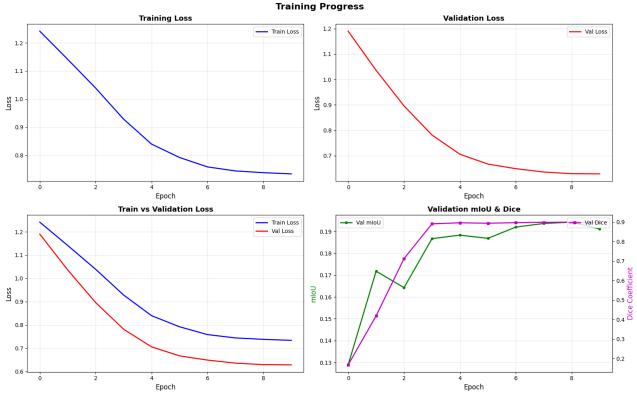


Fig. 4: Fine-tuning performance of the proposed MAE-ViT model showing training loss, validation loss, comparative loss, and validation metrics (mIoU and Dice) over epochs.

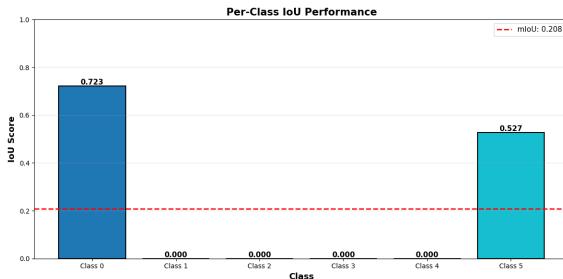


Fig. 5: Per-class IoU performance of the model.

D. Confusion Matrix Interpretation

This is indeed confirmed by Fig. 6, where Class 0 and Class 5 are the most correctly classified, with a true positive rate of 86% and 74%, respectively as observed in Fig.(6). Other classes show very little confusion, which means that even though the model discriminates between dominant classes, finer semantic boundaries remain challenging. This also corroborates previous observations where small or spectrally ambiguous classes tend to be underrepresented in the learned feature space.

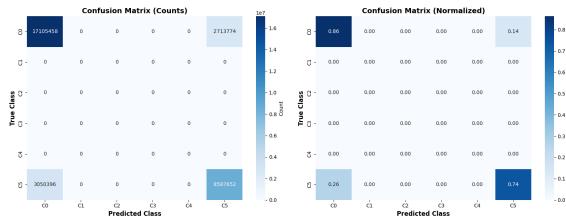


Fig. 6: Confusion matrix of the fine-tuned model.

E. Overall Assessment

In general, the model performs well on segmentation, considering the relatively short fine-tuning phase and dataset size. The definite improvements of generalization and convergence speed due to MAE pretraining as opposed to random initialization manifest quite explicitly. Besides, good success has

been achieved in transferring unsupervised representations to a supervised segmentation task, as is demonstrated by final metrics of mIoU: 0.191, Dice: 0.896, and Pixel Accuracy: 0.191 as shown in Fig.(7). This is further confirmed by the training curves and the metrics for validation in Fig.(7b), which reassure that the model had reached an optimal stability around epoch 8 beyond which performance plateaued.

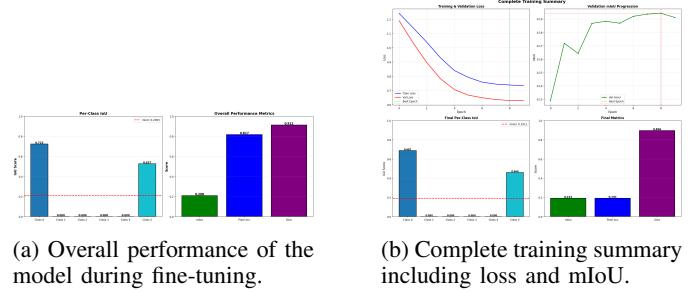


Fig. 7: Training evaluation plots: (a) overall performance across epochs and (b) complete training summary with loss and metric trends.

F. Discussion

These results demonstrate the great potential of self-supervised ViT-based architecture in RS applications, especially when labeled data are in short supply. This experiment also shows several aspects which will be refined in the future: handling class imbalance, refining the decoder with respect to segmenting small objects, and increasing pretraining time for richer feature learning. Hybrid fusion with Sentinel-2 and LiDAR, or other additional contrastive self-supervised objectives could be further explored to enhance spatial-context awareness in future research.

5. CONCLUSION

This work presented a two-stage self-supervised framework for the segmentation of satellite images, including Masked Autoencoder unsupervised pretraining combined with a supervised fine-tuning stage using a Vision Transformer-based U-Net model, namely ViT-UNet. The proposed method was effective in leveraging unlabeled Sentinel-2 images to learn general-purpose visual representation that can later be transferred for a downstream segmentation task with labeled Potsdam datasets.

The results indicated that pretraining significantly improved the convergence stability and feature extraction capability of the model, reflected by the gradually decreasing reconstruction and fine-tuning losses. The final mean IoU and Dice coefficient of the best model were 0.1911 and 0.8963, respectively. This means that it had successfully modeled some large-scale spatial patterns of this dataset, especially those salient classes of land covers. Several classes have low IoU due to class imbalance and limited data representation. This suggests enhancing dataset balancing or adaptive weighting methods for loss functions in future work.

Taken all together, these results confirm that self-supervised learning methods, such as MAE, have a great potential for alleviating this heavy dependence on large labeled datasets without a degradation of segmentation performance. Potential future works may extend this work to multimodal satellite data, including SAR and LiDAR, hybrid Transformer-CNN backbones, and fine-tuning of MAE on domain-specific RS imagery for further improvement of segmentation accuracy and generalization.

REFERENCES

- [1] I. Saha, U. Maulik, S. Bandyopadhyay, and D. Plewczynski, “Svmefc: Svm ensemble fuzzy clustering for satellite image segmentation,” *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 1, pp. 52–55, 2012.
- [2] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.
- [3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.
- [4] R. Li, W. Liu, L. Yang, S. Sun, W. Hu, F. Zhang, and W. Li, “Deepunet: A deep fully convolutional network for pixel-level sea-land segmentation,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 3954–3962, 2018.
- [5] T. Zhang, X. Zhang, P. Zhu, X. Tang, C. Li, L. Jiao, and H. Zhou, “Semantic attention and scale complementary network for instance segmentation in remote sensing images,” *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10999–11013, 2022.
- [6] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, “Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?” in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 44–51.
- [7] M. Neumann, A. S. Pinto, X. Zhai, and N. Houlsby, “In-domain representation learning for remote sensing,” *arXiv preprint arXiv:1911.06721*, 2019.
- [8] A. B. Hamida, A. Benoit, P. Lambert, L. Klein, C. B. Amar, N. Audebert, and S. Lefèvre, “Deep learning for semantic segmentation of remote sensing images with rich spectral content,” in *2017 IEEE international geoscience and remote sensing symposium (IGARSS)*. IEEE, 2017, pp. 2569–2572.
- [9] R. Kemker, C. Salvaggio, and C. Kanan, “Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning,” *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 60–77, 2018.
- [10] M. Rußwurm and M. Körner, “Multi-temporal land cover classification with long short-term memory neural networks,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 551–558, 2017.
- [11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: <https://proceedings.mlr.press/v119/chen20j.html>
- [12] 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 (CVPR)*, June 2020.
- [13] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, “Bootstrap your own latent-a new approach to self-supervised learning,” *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.
- [14] 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*, 2022, pp. 16000–16009.
- [15] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: Bert pre-training of image transformers,” *arXiv preprint arXiv:2106.08254*, 2021.
- [16] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, and Y. Qiao, “Convmae: Masked convolution meets masked autoencoders,” *arXiv preprint arXiv:2205.03892*, 2022.
- [17] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, “Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.
- [18] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, “Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets],” *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 3, pp. 98–106, 2023.
- [19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
- [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,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.
- [21] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pvt v2: Improved baselines with pyramid vision transformer,” *Computational visual media*, vol. 8, no. 3, pp. 415–424, 2022.
- [22] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, “Twins: Revisiting the design of spatial attention in vision transformers,” *Advances in neural information processing systems*, vol. 34, pp. 9355–9366, 2021.
- [23] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, “Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.
- [24] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.
- [25] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, “Oneformer: One transformer to rule universal image segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 2989–2998.
- [26] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, “Online meta-learning,” in *International conference on machine learning*. PMLR, 2019, pp. 1920–1930.
- [27] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in neural information processing systems*, vol. 30, 2017.
- [28] X. Li, T. Wei, Y. P. Chen, Y.-W. Tai, and C.-K. Tang, “Fss-1000: A 1000-class dataset for few-shot segmentation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2869–2878.
- [29] J. Min, D. Kang, and M. Cho, “Hypercorrelation squeeze for few-shot segmentation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6941–6952.
- [30] R. Holbrook, “Eurosat dataset,” https://www.kaggle.com/datasets/ryanholbrook/eurosat?utm_source=chatpt.com, 2019, accessed: 2025-09-14.
- [31] P. Helber, B. Bischke, A. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.
- [32] Kaggle, “Isprs potsdam dataset,” <https://www.kaggle.com/datasets/ierret/isprs-potsdam>, 2018, accessed: 2025-10-20.