MoCo-Enhanced TransU-Net: A Self-Supervised Learning Approach for Landslide Prediction Using Unmanned Aerial Vehicle (UAV) Images

Wun Puo Lim 1, a), Shih Yin Ooi 1, 2, b), Yee Jian Chew 1, 2, c), Yoshitaka Sakata 3, d) and Janjit Iamchaturapatr 4, e)

*1Faculty of Information Science and Technology (FIST), Multimedia University, Jalan Ayer Keroh Lama, Melaka, 75450, Malaysia*

*2Centre for Advanced Analytics (CAA), COE for Artificial Intelligence, Multimedia University, Jalan Ayer Keroh Lama, Melaka, 75450, Malaysia*

*3Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Lizuka-city, Fukuoka, 820-8502, Japan*

*4King Mongkut’s University of Technology North Bangkok, Bang Sue, Bangkok, Thailand*

*b) Corresponding author: syooi@mmu.edu.my*

*a) lim.wun.puo@student.mmu.edu.my*

*c)chewyeejian@mmu.edu.my*

*d)sakata.yoshitaka414@mail.kyutech.jp*

*e) janjit.i@cit.kmutnb.ac.th*

**Abstract.** In recent years, the number of landslide cases has increased, posing a threat to both infrastructure and human safety. Traditional landslide detection methods require a significant amount of manpower, materials, and financial resources. Therefore, governments or developers may be reluctant to conduct monitoring in remote or undeveloped areas. Unmanned Aerial Vehicles (UAVs) have emerged as a powerful tool for collecting land data, and deep learning (DL) models are increasingly used for landslide detection. However, these models require labeled datasets and a significant amount of time to produce. Therefore, this study proposes a novel deep learning framework that integrates Momentum Contrast (MoCo), a self-supervised learning technique, with TransU-Net (a hybrid CNN-Transformer model) to enhance landslide prediction using UAV imagery from the Berembun Forest Reserve. MoCo is used to pre-train a ResNet-50 encoder on unlabeled 512×512-pixel images using contrastive learning. This aims to enhance feature representations without relying on manual annotations. The MoCo pre-trained weights were transferred into the encoder of TransU-Net and used for fine-tuning semantic segmentation on 256×256 labeled image patches. This study compared TransU-Net with and without the MoCo augmentation. TransU-Net without MoCo had the lowest performance, with results of F1-Score 84%, Precision 85%, Recall 83%, and Intersection over Union 74%. For TransU-Net with MoCo, the results achieved are F1 Score 91%, Precision 92%, Recall 91%, and Intersection over Union 85%. These results demonstrate that integrating MoCo enhances segmentation accuracy and efficiency, and it also provides a practical solution for landslide detection in unlabeled data.

# Introduction

Landslides can be referred to as landslips, rockslides, mudslides, or earthslides. A landslide is the collapse of rocks under the action of gravity. They often occur in steep areas of mountainous areas. When landslides occur, loud noises, smoke, and dust are emitted, and the rocks quickly disintegrate and collapse at the lower part. Several factors have contributed to the landslide, including heavy rain, volcanic activity, earthquakes, and mining [1]. Therefore, landslide detection is essential, and traditional methods, such as on-site manual surveys, ground measurements, and hydrological monitoring, require considerable manpower, materials, and financial resources. As an alternative, Unmanned Aerial Vehicles (UAVs) have gained popularity due to their ability to capture high-resolution images of rugged terrain quickly [2].

Additionally, UAV images can be processed using deep learning (DL) to automate landslide prediction. However, deep learning models are facing several real-world limitations. First, the models most rely on fully supervised learning and require large volumes of labeled data. This limitation has resulted in increased time consumption for producing challenging terrains and has become costly [3][4]. Second, the landslide dataset suffers from a class imbalance problem, where the number of landslide pixels is less than that of background pixels. This imbalance makes it challenging for the model to learn meaningful representations of landslide areas, often resulting in poor boundary detection or missed segments [5]. Furthermore, [6] observed a decline in performance because of data imbalance. Third, the model’s performance often drops when it is exposed to shadows, thick vegetation, or changing lighting conditions [6][7]. Therefore, [6] designed a model that incorporates elements of CNNs and Transformers to perform effectively in these challenging scenarios.

To address these limitations, we have proposed a novel model that combines Momentum Contrast (MoCo) with TransU-Net. TransU-Net is a hybrid Convolutional Neural Network (CNN)–Transformer architecture, and Moco is a self-supervised learning technique. Additionally, Moco can learn from unlabeled drone imagery through contrastive learning, thereby reducing its reliance on manual labeling. In this model, ResNet-50 is used in MoCo pre-training to capture meaningful representations from unlabeled 512×512-pixel high-resolution UAV images captured in the Berembun Forest Reserve. ResNet-50 was chosen because it strikes a balance between depth and efficiency, and its residual structure makes it well-suited for extracting complex visual patterns from UAV images. After that, the pre-trained weights were transferred to the encoder of TransU-Net and fine-tuned for semantic segmentation on labeled 256×256-pixel images of the UAVs. The use of smaller patches in this stage is based on GPU memory limitations (8 GB VRAM), which caused runtime errors during training with 512×512 pixels images. This adjustment ensures training feasibility while maintaining sufficient detail for segmentation. In summary, our approach aims to enhance the accuracy of landslide prediction and mitigate deployment challenges in realistic remote sensing environments.

# Literature review

Deep learning (DL) has become a powerful approach for landslide detection using remote sensing or drone imagery. It is common for existing techniques to employ encoder and decoder structures that maintain spatial understanding by incorporating skip connections. Even though these models generally perform well in semantic segmentation tasks, they often struggle to maintain accuracy when detecting small or irregularly aped landslides [3]. To overcome these limitations, recent models have developed innovative construction methods. For instance, [8] developed SCGC-Net to enhance landslide segmentation across various terrain types, achieving a precision of 88.60%, recall of 85.70%, IoU of 73.53%, and F1 Score of 84.75%. This method achieves better segmentation in complex terrain but still requires labeled datasets for training. Moreover, [9] employed Mask R-CNN for post-earthquake landslide detection, achieving a precision of 93.28%, a recall of 89.74%, and an F1-score of 90.25%. This model is an improved version for fast seismic landslide detection have strong boundary localization. However, this model incurs a computational cost, and subtle terrain transitions may still be missed due to its reliance on object proposals. Furthermore, [4] and [10] have studied the TransU-Net model for landslide detection. TransU-Net is a transformer-based model that integrates a CNN encoder with a Vision Transformer (ViT) module to exploit both local features and global context. [4] looked at how various semantic segmentation models, called TransU-Net, U-Net, and U-Net 3+, were utilized for the long-term mapping of landslides. It was found that TransU-Net did better than WideResNet, but its performance fell behind that of U-Net 3+ in terms of long-term accuracy and resilience. On the other hand, [10] investigated the application of the TransU-Net model for landslide classification in the Laonong River Basin, southern Taiwan. When using this two-class labeling scheme, TransU-Net performed best, delivering a precision of 91.66%, a Recall of 83.2%, and an F1-score of 87.2%. At the same time, like supervised models, its success is largely influenced by having adequate unlabeled data and sufficient training data. Additionally, TransU-Net faces some constraints due to the use of much faster and more advanced computing systems.

To reduce reliance on labeled data, self-supervised learning (SSL) has garnered significant attention. [11] proposed self-supervised learning (SSL) method, Momentum Contrastive Learning (MoCo), which is a contrastive learning method that can be pre-trained on unlabeled images. Two distinct augmented views are created from the image and labeled as query and key, and each is passed through its respective encoder. Backpropagation modifies the query encoder, whereas the key encoder is updated using a moving average, ensuring its features remain stable. It utilizes InfoNCE loss as its contrastive loss, as shown in Equation (1) [12]:

(1)

From this formula, *q* represents the output of the query encoder, *k+* is the positive key, and represents the negative samples from the queue. Lastly, adjusts the strength of the similarity values assigned to detectives. This equation aims to make the model show higher similarity to the query’s positive matches and less similarity to all the other negative keys. Additionally, the numerator represents the score of the match between the query and the correct results. In contrast, the denominator represents the sum of the scores of all results, including both correct and incorrect ones. Additionally, the goal of the loss function is to guide queries towards similar instances while moving away from those that are dissimilar. MoCo works differently from many other contrastive learning frameworks, as it employs a dynamic set of negative samples rather than relying on large batches. To prevent the model from relying on new negative samples, MoCo places past negative sample keys in a queue and uses them instead of increasing the overall batch size.

Additionally, MoCo maintains a dynamic queue of negative samples, which enables the storage of past representations without requiring a large batch size. Additionally, MoCo works effectively on all GPU sizes without requiring the processing of large data batches. To manage memory effectively, a queue is used during training to ensure there are many negative examples. Due to the momentum encoder, MoCo can maintain consistent learning and protect against fluctuations in the representations [13]. Moreover, [3] has highlighted the potential of MoCo for geospatial applications. From their findings, MoCo may be used to simplify manual annotations and improve the stability of models, particularly when the data is not properly balanced. Based on a few studies, MoCo has several advantages, but it still has some limitations. MoCo requires a large pool of diverse negative samples for effective learning, which can limit its performance if the negatives are insufficient. Additionally, its dual-encoder setup and dynamic queue add complexity, making it more challenging to implement compared to simpler contrastive learning methods [13]. In summary, existing models, whether based on CNNs, Transformers, or domain-specific models, all share some common limitations, including reliance on labeled datasets, difficulty in addressing class imbalance, and limited robustness under varying environmental conditions.

To overcome the limitations of previous models, this study integrates Momentum Contrast (MoCo) with the TransU-Net model. A ResNet-50 encoder is first pre-trained using MoCo on unlabelled UAV images to learn robust feature representations. This pre-trained encoder will be fine-tuned with the TransU-Net framework on unlabeled image patches for landslide segmentation. This combined approach reduces reliance on manual annotations, improves performance under class imbalance, and enhances model generalization in complex environmental conditions.

# Methodology

In this study, we propose a framework that integrates a self-supervised learning technique with the TansU-Net model for landslide segmentation from UAV images. This framework involves five stages: the dataset acquisition stage, data preprocessing stage, MoCo pre-training stage, TransU-Net Training stage, and the testing stage. Additionally, to benchmark the effectiveness of MoCo, we prepared two experimental setups: one with the TransU-Net model using MoCo and Another without MoCo. In the data acquisition stage, the UAV dataset used in this experiment was obtained from the Berembun Forest Reserve in Malaysia, presented by [14]. This dataset comprises 15 high-resolution UAV images with a resolution of 5472 × 3648 pixels, captured using a DJI Phantom 4 RTK drone equipped with multispectral and RGB sensors. Since models like MoCo and TransU-Net require fixed input sizes, the original images are split into non-overlapping patches of 512 × 512 pixels (resulting in a total of 2,338 images) for the MoCo pre-training stage and 256 × 256 pixels (resulting in a total of 9,822 images) for the TransU-Net model training stage. The images are unlabelled. The process of splitting the original images into 256 × 256 pixels also generated mask images. The mask images were prepared for TransU-Net training because the model required mask images during training. Due to the device’s limited performance (8GB VRAM), the TransU-Net model was unable to complete the experiment using 512x512 and 384x384 pixel UAV images.

Next, in the MoCo pre-training stage, a ResNet-50 encoder [15] was used to train unlabeled UAV images. ResNet-50 was selected due to its strong balance between depth and feature representation capability, and it is a common choice for MoCo. Moreover, in MoCo training, a queue size of 4096 was used during MoCo pre-training to ensure there were many different negative samples in the process. Furthermore, the temperature parameter (τ) was set to 0.1 to ensure that the logits better separated the positive and negative samples. Additionally, a learning rate of 0.0003 was used to optimize stochastic gradient descent, and this scheme was paired with cosine annealing to bring its progress to a smooth close. The model was trained 150 times on a batch size of 32 to keep within the GPU memory limits. To determine the optimal number of training epochs, five experiments were conducted, and the results showed that the performance was lower than that at 150 epochs until 200 epochs. Therefore, in this experiment, 150 epochs were selected as the optimal training duration. After MoCo pre-training, the pre-trained ResNet-50 encoder weights are saved for later use in TransU-Net training.

After obtaining the MoCo-pre-trained ResNet-50 encoder weights, these weights were used to initialize the TransU-Net encoder. There are two versions of TransU-Net were trained. The first version was TransU-Net without MoCo. Another version integrated MoCo-pretrained weights with TransU-Net, consisting of the MoCo-pretrained weights as the encoder, a Visual Transformer (ViT) module for modeling long-range spatial dependencies, and a decoder with skip connections and bilinear up-sampling for reconstructing spatial details. Then, the 1×1 convolutional layer was used to generate pixel-level binary segmentation maps. During the two versions of the TransU-Net training stage, 60% of the 256 × 256 image patches were used to train the model, employing the Adam optimizer with a learning rate of 0.0001 for the training. This was done to avoid the significant change in weights that could lead to confusion in the pre-trained encoder. Figure 1 shows the workflow of the MoCo-Enhanced TransU-Net for landslide prediction using UAV images.

A diagram of a computer flowchart

AI-generated content may be incorrect.

**Figure 1.** Workflow of the MoCo-Enhanced TransU-Net for landslide prediction using UAV images

Additionally, to ensure fair evaluation, the TransU-Net model with and without MoCo is trained and tested using the same dataset (patch size of 256×256). Because the GPU had only 8 GB of memory, the batch size was fixed at 4, and training proceeded for 150 epochs. The 150 epochs chosen for training were selected to strike a balance between sufficient learning and avoiding overfitting. The initial training was conducted with a small number of epochs, typically 50 to 100. However, experiments showed that although the model converged quickly, extending the training time continued to slightly improve F1 performance while reducing the variance of the test loss. However, after 200 epochs, we observed diminishing returns on performance improvement and a slight increase in the volatility of the test loss, which indicated overfitting. Therefore, 150 epochs were chosen as the optimal point to balance learning ability and generalization ability. Finally, the trained TransU-Net model was saved into a Python pickle file.

For the last stage, testing, 40% of the 256 × 256-pixel images were used to evaluate its performance. The evaluation produced results including the F1-score, precision, recall, Intersection over Union (IoU), and test loss. In conclusion, by leveraging a combination of MoCo pre-training, a Vision Transformer for global attention, and an efficient decoder for segmentation, the TransU-Net model effectively learns to detect landslides from drone images with high accuracy and spatial precision. This ensures robustness by regularly saving model checkpoints at specific intervals, preventing data loss and allowing for future fine-tuning. Additionally, during the pre-training stage, 512 × 512-pixel images are utilized to help MoCo capture more diverse and global visual patterns, thereby enhancing the encoder’s ability to learn generalizable features. Then, in the TransU-Net training stage, 256 × 256 pixels images are used to balance spatial detail with computational efficiency, enabling precise pixel-level segmentation. Although both image sizes come from the same original UAV dataset, they are processed differently to serve the needs of each learning phase.

# eXPERIMENT RESULT

This section evaluates the performance of TransU-Net with and without the MoCo model. Additionally, these two models are compared to several deep learning approaches. The evaluation focuses on four standard segmentation metrics: Precision, Recall, F1-score, and Intersection over Union (IoU). These metrics are widely used in remote sensing and semantic segmentation tasks [3][6][9]. Precision (see Equation (2)) measures the accuracy of the predicted landslide pixels. In landslide detection, high precision means the model minimizes false alarms, such as bare soil or shadows, that are mistaken for landslides. Next, recall represents the number of actual landslide pixels that the model successfully detected.

In landslide scenarios, a high recall is critical because missing true landslide areas (false negatives) can result in serious risks in hazard-prone areas being overlooked. Equation (3) shows the formula for recall. In addition, the F1-score can be used to evaluate the accuracy of positive predictions and the detection of actual positives (refer to Equation (4)). This is particularly useful in landslide detection, as datasets are often unbalanced, and landslides only occupy a small portion of the image. Lastly, Intersection over Union (IoU), shown in Equation (5), measures the overlap between the predicted landslide area and the actual ground truth. In landslide segmentation, a higher IoU indicates that the model not only detects the landslide area but also accurately locates its shape and extent. Where all the TP, FP, and FN in these four metrics were represented as true positives (TP), false positives (FP), and false negatives (FN), these metrics are shown below:

(2)

(3)

(4)

(5)

In this experiment, 150 epochs were used because they yielded the best performance at that point. Figure 2 illustrates the training and testing performance of our proposed TransU-Net model with MoCo for 150 epochs. At the beginning of the experiment, the training loss value drops dramatically from epoch 1 to epoch 37, from 0.55 to 0.10. After epoch 37, the train loss gradually decreases to 0.01. For the test loss, it fluctuates slightly but also exhibits a downward trend, eventually stabilizing at around 0.2. This stability shows that the model can generalize well to unlabelled data. Lastly, the F1-score rises steadily from around 0.73 to over 0.9 at around epoch 50 and remains high for the rest of the training epochs with minimal fluctuations. This trend confirms that the model can learn well.

Furthermore, as shown in Table 1, the proposed TransU-Net model enhanced with MoCo achieved the highest F1-score of 91.00%, outperforming the baseline TransU-Net without MoCo (F1-score 84.00%) by a margin of 7.00%. This highlights the benefit of incorporating contrastive pre-training to improve feature representation and segmentation accuracy for landslide detection. Compared to other recent methods, the proposed model also surpasses Mask R-CNN (F1-score 90.25%), SCGC-Net (84.75%), and TransU-Net (ACRS 2024) (87.21%), while maintaining high precision, recall, and IoU. To ensure a fair comparison, the TransU-Net with MoCo without MoCo was trained and evaluated using the same dataset, preprocessing steps, patch sizes, and experimental settings.

A graph with a line graph

AI-generated content may be incorrect.

**Figure 2.** Training and testing performance of the TransU-Net with MoCo for 150 epochs

|  |  |  |
| --- | --- | --- |
| **TABLE 1.** Comparison of performance results between the MoCo model, the model without MoCo, and other approaches to landslide detection models | | |
| **Model** | **With/Without MoCo** | **F1-Score Precision Recall IoU** |
| TransU-Net (Ours)  TransU-Net (Ours)  Mask R-CNN [9] | With MoCo  Without MoCo  Without MoCo | 91.00% 92.00% 91.00% 85.00%  84.00% 85.00% 83.00% 74.00%  90.25% 93.28% 89.74% - |
| SCGC-Net [8]  TransU-Net [10] | Without MoCo  Without MoCo | 84.75% 88.60% 85.70% 73.53%  87.21% 91.66% 83.18% - |

# Limitations and Future Work

This study proposed a deep learning framework that integrates Momentum Contrast (MoCo) with TransU-Net for landslide prediction using UAV images. The experimental results demonstrate promising performance improvements; however, several limitations exist. First, this experiment utilizes only a single dataset, which may limit its generalizability to other geographic locations or imaging conditions. Even the results show that robust feature learning can be achieved through MoCo; however, future work should explore cross-dataset validation to evaluate its transferability better. Second, in this study, an ablation study was performed to examine the effect of MoCo on TransU-Net; however, some limitations prevent us from conducting a more comprehensive sensitivity analysis on training parameters, such as learning rate and temperature τ. Therefore, future work could perform a detailed exploration of the impact of these hyperparameters, providing further insights into the optimal configuration of contrastive MoCo pre-training in geospatial applications. Moreover, due to memory limitations, we chose a 256×256 pixels image for TransU-Net training. For future work, we can explore more pixel images to train on TransU-Net with the MoCo model, comparing the performance achieved with each image size. Lastly, this study focuses on MoCo as the self-supervised learning (SSL) approach. Still, future work can compare its performance with other self-supervised learning (SSL) approaches, such as SimCLR, BYOL, or SwAV, to provide a broader understanding of which frameworks are most effective for landslide detection.

# CONCLUSION

In this study, we explored how combining self-supervised learning with deep learning can improve landslide prediction using UAV images. By using Momentum Contrast (MoCo) to pre-train the model on unlabelled data, we provided the TransU-Net encoder with a strong foundation before fine-tuning it with limited labeled samples. Furthermore, mask images are still needed during TranU-Net training. Because MoCo improves the training process by helping the model learn faster, avoid overfitting on datasets, and generalize better to new or complex terrains, the TransU-Net model still requires mask images for training. The MoCo-enhanced version outperformed TransU-Net and other existing methods on all primary evaluation metrics. It demonstrates that self-supervised learning can be applied in remote sensing, where collecting labeled data is challenging. In the future, this approach could be extended to include various types of data or utilized in real-time systems to support early warnings and landslide prevention.

# Acknowledgments

This research was funded by a Matching Grant from Multimedia University and the King Mongkut’s University of Technology North Bangkok (MMUI/240102).

# References

1. A. Lees, “What Is a Landslide? Types, Causes, Effects & Prevention,” (n.d.).
2. I. Sandric, Z. Chitu, V. Ilinca, and R. Irimia, “Using high-resolution UAV imagery and artificial intelligence to detect and map landslide cracks automatically,” Landslides **21**(10), 2535–2543 (2024). https://doi.org/10.1007/s10346-024-02295-9
3. T.M. Pham, N. Do, H.T.T. Pham, H.T. Bui, T.T. Do, and M.V. Hoang, “CResU-Net: a method for landslide mapping using deep learning,” Mach. Learn.: Sci. Technol. **5**(3), 035008 (2024). https://doi.org/10.1088/2632-2153/ad5f17
4. Z. Hu, B. Yi, H. Li, C. Zhong, P. Gao, J. Chen, Q. Yao, and H. Guo, “Comparative Evaluation of State-of-the-Art Semantic Segmentation Networks for Long-Term Landslide Map Production,” Sensors **23**(22), 9041 (2023). https://doi.org/10.3390/s23229041
5. G. Li, K. Li, G. Zhang, K. Pan, Y. Ding, Z. Wang, C. Fu, and Z. Zhu, “A landslide area segmentation method based on an improved UNet,” Sci Rep **15**(1), 11852 (2025). https://doi.org/10.1038/s41598-025-94039-5
6. L. Wu, R. Liu, N. Ju, A. Zhang, J. Gou, G. He, and Y. Lei, “Landslide mapping based on a hybrid CNN-transformer network and deep transfer learning using remote sensing images with topographic and spectral features,” International Journal of Applied Earth Observation and Geoinformation **126**, 103612 (2024). https://doi.org/10.1016/j.jag.2023.103612
7. X. Li, T. Chen, G. Liu, J. Dou, R. Niu, and A. Plaza, “Joint Grid-Based Attention and Multilevel Feature Fusion for Landslide Recognition,” IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing **17**, 19911–19922 (2024). https://doi.org/10.1109/JSTARS.2024.3491216
8. Y. Fan, P. Ma, Q. Hu, G. Liu, Z. Guo, Y. Tang, F. Wu, and H. Zhang, “SCGC-Net: Spatial Context-Guided Calibration Network for Multisource RSI Landslides Detection,” IEEE Trans. Geosci. Remote Sensing **63**, 1–17 (2025).
9. R. Fu, J. He, G. Liu, W. Li, J. Mao, M. He, and Y. Lin, “Fast Seismic Landslide Detection Based on Improved Mask R-CNN,” Remote Sensing **14**(16), 3928 (2022).
10. S.-N. Tsai and F. Tsai, “Exploring the capability of TransUNet model for landslide classification,” ACRS 2024, (2024).
11. K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum Contrast for Unsupervised Visual Representation Learning,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, Seattle, WA, USA, 2020), pp. 9726–9735.
12. N. Bendib, “Paper explained — Momentum Contrast for Unsupervised Visual Representation Learning [MoCo]” Medium, (2022).
13. H. Amit, “Easily Explained: Momentum Contrast for Unsupervised Visual Representation Learning,” Medium, (2024).
14. Y.J. Chew, S.Y. Ooi, S. Mohd-Razali, Y.H. Pang, and Z. You Lim, “Enhancing Land Management through U-Net Deep Learning: A Case Study on Climate-Related Land Degradation in Berembun Forest Reserve in Malaysia,” JOIV: Int. J. Inform. Visualization **8**(4), 2419 (2024).
15. J.-R. Lee, K.-W. Ng, and Y.-J. Yoong, “Face and Facial Expressions Recognition System for Blind People Using ResNet50 Architecture and CNN,” Journal **2**(2), 284–298 (2023).