A MULTI-SCALE VISION TRANSFORMER-BASED MULTIMODAL GEOAI MODEL FOR MAPPING ARCTIC PERMAFROST THAW

Wenwen Li, Chia-Yu Hsu, Sizhe Wang, Zhining Gu
School of Geographical Sciences and Urban Planning
Arizona State University
Tempe, AZ, USA
{wenwen, chsu53, wsizhe, zhiningg}@asu.edu

Yili Yang, Brendan M. Rogers, Anna Liljedahl
Woodwell Climate Research Center
Falmouth, MA, USA
{yyang, brogers, aliljedahl}@woodwellclimate.org

ABSTRACT

Retrogressive Thaw Slumps (RTS) in Arctic regions are distinct permafrost landforms with significant environmental impacts. Mapping these RTS is crucial because their appearance serves as a clear indication of permafrost thaw. However, their small scale compared to other landform features, vague boundaries, and spatiotemporal variation pose significant challenges for accurate detection. In this paper, we employed a state-of-the-art deep learning model, the Cascade Mask R-CNN with a multi-scale vision transformer-based backbone, to delineate RTS features across the Arctic. Two new strategies were introduced to optimize multimodal learning and enhance the model's predictive performance: (1) a feature-level, residual cross-modality attention fusion strategy, which effectively integrates feature maps from multiple modalities to capture complementary information and improve the model's ability to understand complex patterns and relationships within the data; (2) pre-trained unimodal learning followed by multimodal fine-tuning to alleviate high computing demand while achieving strong model performance. Experimental results demonstrated that our approach outperformed existing models adopting data-level fusion, feature-level convolutional fusion, and various attention fusion strategies, providing valuable insights into the efficient utilization of multimodal data for RTS mapping. This research contributes to our understanding of permafrost landforms and their environmental implications.

Keywords Vision transformer · instance segmentation · Artificial Intelligence · Remote Sensing · Multi-modal fusion

1 Introduction

As awareness grows, it becomes evident that the temperature in Arctic regions has been rising at a rate at least twice as fast as the global average increase [5]. Permafrost landscapes are significantly impacted by Arctic warming [21]. Abrupt permafrost thaw and climate shifts (e.g., increasing temperature and precipitation) have caused active layer detachment or slope failure, expediting the formation of distinctive landforms known as Retrogressive Thaw Slumps (RTSs) [35, 17, 40]. The progression of RTSs exposes underlying soil that contains abundant organic materials, increasing the risk of accumulation of heavy metal chemicals [18, 17]. Permafrost thaw is damaging vegetation cover and disrupting hydrological and ecological processes in surrounding areas [45]. Both the ecological structure and infrastructure are vulnerable to the impacts of permafrost thaw. Additionally, the intensification of carbon emissions is another rising concern associated with this process [45]. As a result, understanding RTS dynamics becomes pivotal for understanding permafrost thaw and its impacts on terrestrial alterations, local ecosystems, the hydrological system, and the carbon cycle [45, 52].

Mapping RTSs offers insights into the spatiotemporal changes in the permafrost landscape and local environmental dynamics. However, no pan-Arctic maps of RTS exist, hindering our ability to assess their full distribution, changes over time, and impacts on ecosystems and the carbon cycle. Traditionally, RTS mapping relied heavily on manual digitization using the Esri ArcGIS platform [22, 37]. However, this method is very time-consuming, and is limited to small-area research [17, 52]. Recent advancements in Geospatial Artificial Intelligence (GeoAI) [24] and Deep Learning (DL) have led to new models demonstrating outstanding performance in various computer vision tasks, including object detection [2, 8, 19, 15, 1, 26], and semantic and instance segmentation [12, 2, 27]. Specifically, semantic segmentation classifies every pixel in an image into predefined categories without differentiating individual objects of the same class. Object detection, on the other hand, identifies and localizes objects by predicting bounding boxes around them along with their associated class labels. Instance segmentation combines these goals by classifying each pixel while distinguishing between distinct instances of the same object class.

These cutting-edge GeoAI techniques have been employed to support diverse geographical applications, including plane and airport detection from aerial images [47], land cover segmentation [42], flood mapping [23], and crop yield estimation [49]. In recent years, advanced GeoAI models have started to be applied to enhance the automation capabilities in RTS mapping. Semantic segmentation models, such as U-Net [38], DeepLabV3 [3], and DeepLabV3+ [4], have been frequently employed on remote sensing images to delineate RTSs in the Arctic [35, 52] and on the Tibetan Plateau [18], providing valuable insights into the potential of DL models for RTS research.

While the pioneering research mentioned above has significantly advanced the automated mapping of RTSs, prior studies have typically formulated RTS segmentation as a semantic segmentation problem. Although these methods can identify RTS and non-RTS pixels, they cannot distinguish RTSs as individual objects. Due to spatiotemporal variations in soil moisture, vegetation cover, water content, and terrain, the shapes of RTSs and their rates of change can vary [11, 52]. Additionally, each RTS uniquely impacts its local environment [11]. Therefore, providing fine-grained, instance-level segmentation of RTSs is crucial for tracking their spatial extent, growth patterns, and characteristics over time, thereby enhancing the understanding of permafrost dynamics across heterogeneous Arctic landscapes.

Instance segmentation effectively identifies different objects from the image, which is suitable for recognizing individual RTS features. The process can employ GeoAI and deep learning models like Convolutional Neural Networks (CNNs), which utilize CNN-based encoders such as ResNet [13] and VGG-Net [39]. An alternative approach involves Vision Transformer (ViT)-based feature extractors. Originally developed for natural language processing, Transformers have been adapted for visual tasks through the development of ViT, enhancing their applicability in image segmentation tasks. Various instance segmentation models, adopting either a transformer or a CNN backbone, typically use an encoder followed by a decoder, also known as a segmentation head. Popular models include Mask R-CNN [12] and Cascade Mask R-CNN [2]. Mask R-CNN is a two-stage model which first generates proposals for object regions and then, in the second stage, refines these proposals by classification and regression tasks. Cascade Mask R-CNN makes further improvements to incorporate multiple stages to refine predictions in every stage, which achieves better performance than Mask R-CNN [2].

In addition, multimodal learning has further empowered vision tasks [1, 53]. "Multimodal" refers to the incorporation of various types of data inputs, such as images and text, to enrich the learning process. Each modality contributes unique and complementary information [50], allowing for the generation of important features related to the targets. As a result, the integration of multiple modalities holds the potential to provide comprehensive feature information, serving as prior knowledge for GeoAI models to acquire rich information through joint learning [53]. In the geospatial domain, the integration of different data modalities from different sources provides multifaceted characteristics of landforms. For instance, optical images visually distinguish landforms such as valleys, hills, and rivers. Multi-spectral images record data within specific wavelength bands beyond what is visible to human eyes, enabling in-depth analysis in various fields like agriculture and environmental monitoring. Light Detection and Ranging (LiDAR) uses laser light to measure distances, thereby generating high-resolution elevation models that offer detailed insights into the topography of landforms.

This paper aims to take advantage of multimodal data across the Arctic and cutting-edge deep learning models to achieve automated RTS delineation. This is achieved by the introduction of a new multimodal learning strategy and the utilization of vision-transformer-based multi-scale instance segmentation techniques. The rest of the paper will be organized as follows. Section 2 will introduce relevant work on GeoAI and deep learning models and multimodal data usage in RTS mapping. Section 3 describes study areas for RTS mapping. Section 4 describes our proposed methodology in detail. Section 5 presents the experimental results and analysis. Section 6 concludes our work and proposes future research directions.

2 Related Work

2.1 Deep learning-based RTS mapping

Recent years have witnessed exciting advances in artificial intelligence techniques, particularly in object detection and segmentation tasks within computer vision. The use of segmentation DL models for RTS mapping is becoming a popular and promising solution for large-scale studies [18, 35, 17, 41, 45, 52]. These models have been employed to assess RTS mapping performance, including U-Net [35, 45, 52], U-Net++ [35], DeepLabv3 [35], and DeepLabv3+ [18, 17]. For instance, [35] used DeepLabv3 and U-Net++ to map RTS and achieved an Intersection over Union (IoU) score of 0.58 with U-Net++. [52] expanded the spatial extent of the RTS dataset and enhanced the DEM data by employing relative elevation and enhanced shaded-relief layer. By applying a label sampling approach and label smoothing strategy, the model achieved the highest IoU score of 0.71 with U-Net3+, outperforming the U-Net++, TransU-Net, and ResU-Net models.

All these studies treat RTS delineation as a semantic segmentation task. However, as [37] pointed out, RTS landforms are dynamic, under varying environmental conditions including precipitation, temperature, and stream erosion influencing spatial and temporal variations in physical properties such as soil moisture, vegetation cover, and elevation contrast [52]. For instance, while RTSs in Banks Island are mainly found along lake shores and valley slopes [35], those near the Lena River are largely characterized by dense shrubs along lake shores [35]. Despite these variations, current studies have not focused on mapping and analyzing the unique characteristics of individual RTS features. Fortunately, instance segmentation models offer a solution to address this gap.

Instance segmentation models, such as Mask R-CNN, incorporate Regional Proposal Network (RPN) to generate candidate object proposals based on generated feature maps and achieve the parallel classification, bounding box regression, and mask generation. Using this approach, individual objects, such as RTS, can be segmented. In our paper, we employed an enhanced model of Mask R-CNN, integrating it with a multi-scale vision transformer-based backbone and a multi-stage segmentation to further improve its segmentation performance.

2.2 Multimodal learning in landform mapping

In natural feature detection, the use of multi-modalities ranging from optical image and other data modalities has grown [1, 53, 44]. By fusing heterogeneous data sources, models harness complementary features through joint learning, leading to optimized prediction outcomes [53]. To effectively delineate RTS, [18] combined optical images with DEM and its derivatives (e.g., slope and topographic position index), while [52] incorporated NDVI and other DEM features including relative elevation and enhanced shaded-relief features. While these studies successfully leverage multimodal data to map landform features, the fusion strategies employed are relatively traditional, and a gap remains in integrating cutting-edge multimodal fusion strategies into landform mapping research.

For example, many studies in RTS mapping [18, 35, 45, 52] have employed a data-level fusion approach, in which additional data modalities are integrated as extra channels alongside RGB channels to serve as inputs to DL models. For instance, [52] adapted the U-Net model, which originally takes 3 channels (red, green, and blue) as inputs. They expanded the input by concatenating the RGB image with NDVI, relative elevation, and enhanced shaded relief, resulting in a hybrid 6-channel input.

An alternative approach to multimodal fusion is feature-level fusion. In this method, feature maps of each modality are extracted independently by backbone feature extractors, such as ResNet [13], VGG-Net [39], and Multi-scale ViT [29]. To effectively integrate diverse types of information, different fusion strategies can be employed on multimodal feature maps. For instance, in mapping natural features, [44] employed CNN-based backbones to individually extract feature maps for each modality, such as optical image, DEM, and DEM derivatives. They then introduced an additional convolutional fusion layer to integrate these modalities. Similarly, [36] leveraged the RGB image and depth data to extract relevant features. After concatenating these features, they added a convolutional layer to merge them, enhancing the model's capability for subsequent object detection tasks involving classification and regression.

In addition to convolutional fusion, attention mechanisms are a powerful tool in deep learning. Attention modules can dynamically focus on the most informative parts (e.g., where the "attention" should be) of input data during processing [43]. By assigning varying weights to different elements, attention enables deep learning models to capture complex relationships and dependencies, even across distant features. For images, attention operates by dividing an image into patches or spatial regions, computing attention scores to identify which regions are most relevant for a given task [6, 30]. These scores weigh the contributions of each region, allowing the model to prioritize important features while de-emphasizing less informative areas. Extending the attention mechanism in a multimodal learning framework can further enhance the "attention" capturing capabilities by leveraging relevant information from different modalities to

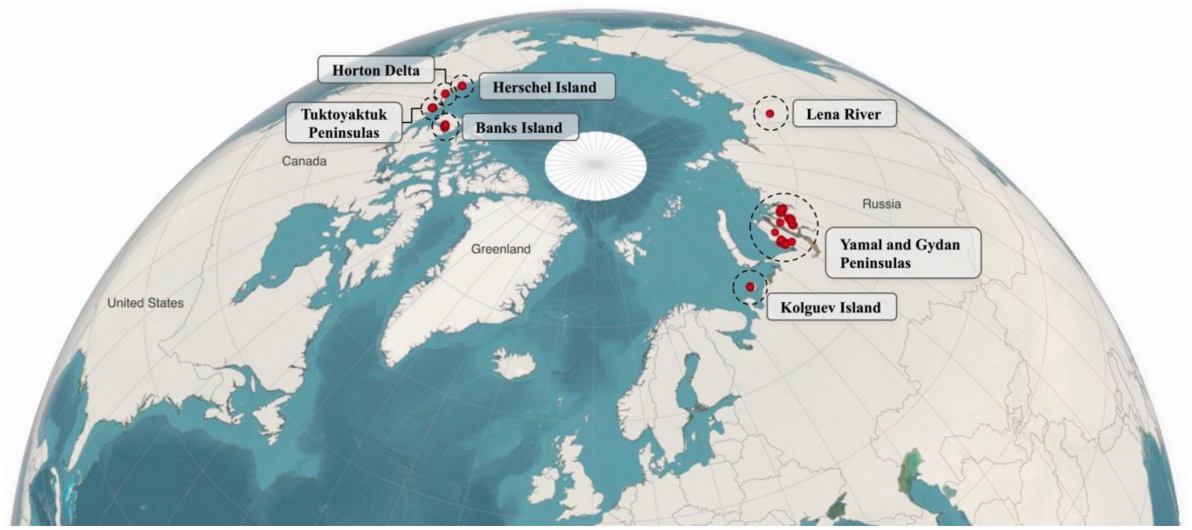


Figure 1: Study Areas. There are 855 RTS samples in total located in seven sites from both Canada and Russia. In Canada, study sites include Herschel Island (41 RTS samples), Horton Delta (43 RTS samples), Tuktoyaktuk peninsulas (165 RTS samples), and Banks Island (174 RTS samples). In Russia, 399 RTSs are from Yamal and Gydan peninsulas, 7 samples are near the Lena River, and the remaining 26 RTS samples are from Kolguev Island.

reinforce the feature extraction process. This approach reinforces important cross-modal features and creates a more informative feature representation, resulting in stronger model prediction capabilities.

Despite its promise, comprehensive research on multimodal fusion strategies regarding their effectiveness, computational efficiency, contributions of data modalities, and the methods' generalizability has been underexplored in landform mapping applications, especially for mapping retrogressive thaw slumps. To address this gap, our research has proposed a novel residual cross-modality fusion strategy that effectively integrates multimodal data with varying characteristics to advance RTS mapping across the heterogeneous Arctic landscape. The next section introduces the study area, followed by a detailed description of our proposed model in Section 4.

3 Study Area

Our study area encompasses seven sites, utilizing data from [52] and [35]. These sites provide a diverse representation of RTSs, covering various environmental and geomorphological conditions. Three sites, including the Yamal and Gydan Peninsulas, the Lena River, and Kolguev Island, are located in Russia. The other four, situated in Canada, comprise Herschel Island, Horton Delta, the Tuktoyaktuk Peninsula, and Banks Island. Figure 1 illustrates the spatial distribution of the RTSs. For the experiment, we collected a total of 855 RTS image scenes, with 399 samples from the Yamal and Gydan Peninsulas, 7 near the Lena River, and 26 from Kolguev Island in Russia. Additionally, 41 samples are from Herschel Island, 43 from Horton Delta, 165 from the Tuktoyaktuk Peninsula, and 174 from Banks Island in Canada.

Each study site presents distinctive environmental conditions for RTSs. For instance, the Yamal and Gydan peninsulas are the only known areas with gas emission craters [55]. In addition, Banks Island and Herschel Island are characterized by extensive ground ice and exhibit the most active RTSs [35]. Herschel Island is further distinguished by shrubby tundra interspersed with small lakes and streams. Horton Delta, in contrast, features steep terrain dominated by the vegetation of dwarf shrub tundra, where RTSs are commonly found on steep slopes. Moreover, Kolguev Island, with its ice-rich permafrost, often sees RTSs on steep coastal bluffs [35].

4 Methodology

To segment RTS features with high accuracy, we implemented our multimodal fusion framework by extending the Cascade Mask R-CNN instance segmentation model [2]. Such an instance segmentation pipeline (Figure 2) starts with the backbone network (the encoder) to extract feature maps from input images as their encoded feature representation. We chose a multi-scale ViT as the model backbone because of its ability to capture global contextual information at

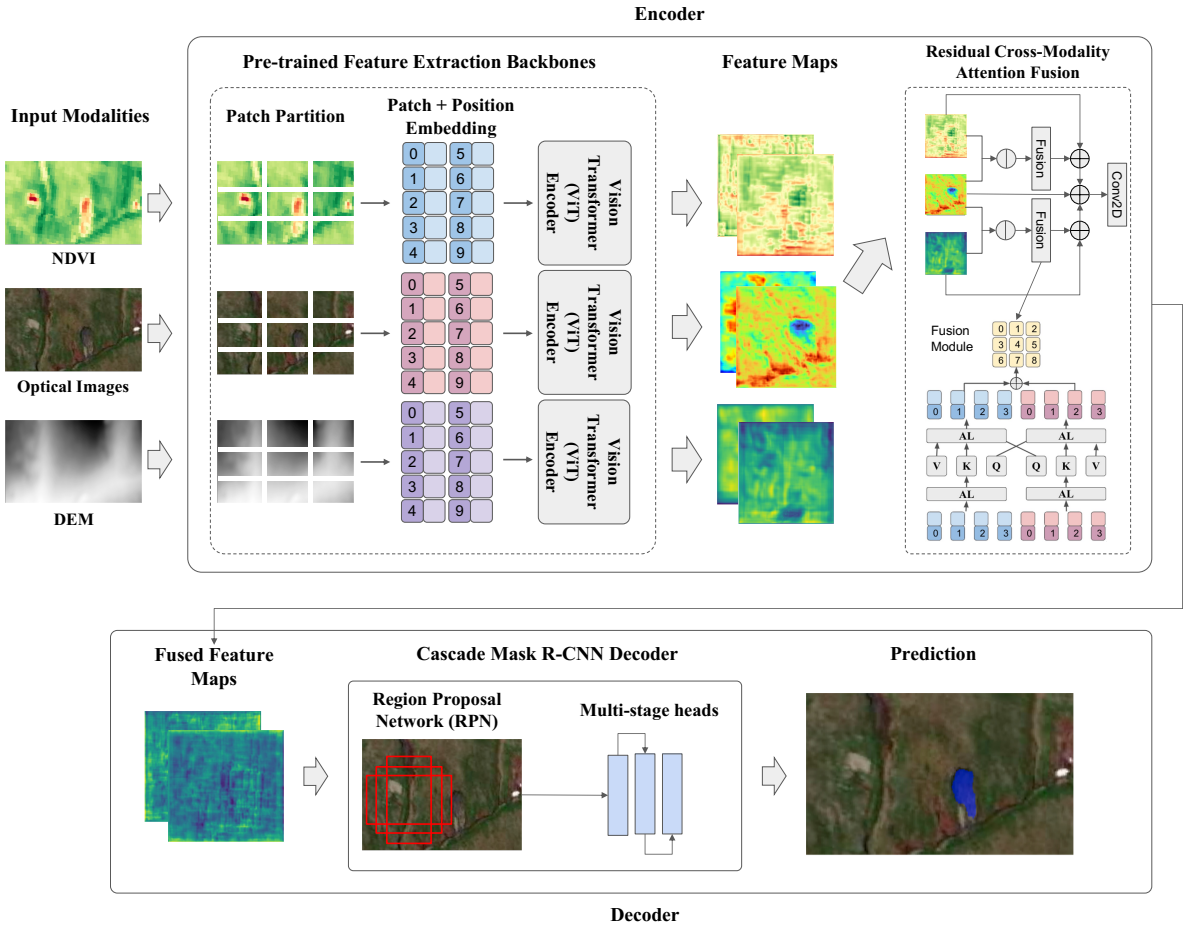


Figure 2: Architecture of multimodal instance segmentation model for RTS mapping.

varying scales and resolutions. The decoder of the instance segmentation model utilizes the resultant feature maps to reconstruct spatial details for accurate instance segmentation.

The aforementioned model was designed to use a single modality as input. As mentioned before, multimodal data provide complementary information to support a better understanding and delineation of RTS features. For example, optical images with RGB bands support the visual inspection of RTS. Since RTS results from landslides and ground subsidence, this process often causes a change in vegetation cover around it. As such, vegetation indices (e.g., NDVI) and spectral bands beyond the visible spectrum, such as Near Infrared (NIR), become critical additional data modalities for capturing vegetation changes in RTS regions relative to their surroundings. Figure 3 illustrates an example of the multimodal data and the corresponding RTS labels created as part of the AI-ready training dataset.

To better accommodate multiple data modalities, we propose two innovative strategies for the image segmentation pipeline: (1) a feature-level residual cross-modality attention fusion and (2) pre-trained unimodal learning followed by multimodal model fine-tuning. Both strategies, as illustrated in Figure 2, focus on modifying the instance segmentation pipeline to enable effective multimodal learning. During pre-trained unimodal learning, each input modality (RGB, NDVI, or NIR) is individually trained on a Cascade Mask R-CNN model with a multi-scale ViT encoder, generating backbone weights optimized for segmentation accuracy. These pre-trained backbone models are then utilized in the multimodal training process. With feature maps for three modalities extracted, the residual cross-modality attention fusion strategy (Figure 2, top right) is applied to generate fused feature maps by combining complementary information across modalities. The resultant feature maps are then sent to the decoder, composed of a Region Proposal Network and multi-stage segmentation heads for instance classification, detection refinement, and final segmentation.

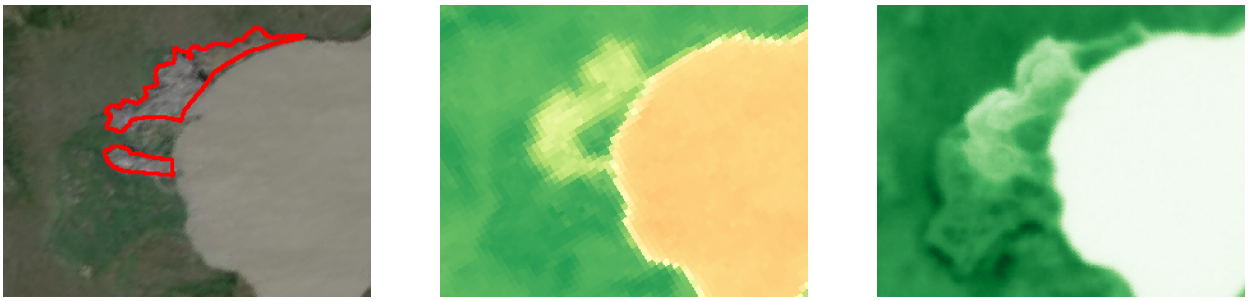
4.1 Multi-scale Vision-Transformer-based feature extraction backbone

In this work, we adopt a multi-scale vision transformer as the backbone model at the pre-training phase. The backbone, also referred to as the encoder, transforms the input image into feature maps. These feature maps are then processed by the segmentation head to reconstruct images for the final predictions in what is also known as the decoding stage. Different from CNN-based backbones which take the entire image as input, ViT divides each image into uniform-sized patches and takes each flattened image patch as input. These flattened patches, each augmented with position embeddings, are then fed into a Transformer Encoder. Multi-scale ViT builds upon the standard ViT by transitioning from a constant resolution to multi-scale resolutions. This is achieved by increasing the number of channels and decreasing the sequence length (or resolution) at various scaled stages [7]. The integration of relative position embeddings and the enhanced pooling connection within the Transformer Encoder significantly improves the model's performance. These enhancements allow the model to consider image semantic information and spatial relationships based on relative locations, which helps reduce computational and memory demands.

4.2 Singular-modal pretraining and multimodal fine-tuning

A challenge for multimodal learning, especially at the feature fusion level, is the memory consumption required for multi-backbones, each dedicated to supporting feature extraction from a single modality. Transformer architectures also often require significantly more memory than traditional CNN-based models due to the complex attention mechanisms they adopt. As a result, introducing multiple feature extraction paths (as shown in Figure 2) typically demands substantial computing resources and GPU capacity, making model training more challenging than in unimodal learning. Furthermore, larger models tend to be more difficult to train and converge. When training data is limited, the model parameters may not be optimally tuned and, therefore, may fail to achieve optimal performance.

To address this challenge, we proposed a phased training strategy, beginning with pre-trained single-modal learning, followed by multimodal model fine-tuning. Each modality's data is first processed through a unimodal instance segmentation pipeline for model pre-training. The backbone weights yielding the best model results for each modality are saved in preparation for subsequent phases of multimodal fine-tuning. Once the best model prediction accuracy is achieved for each modality, the corresponding ViT backbone weights are frozen, making them untrainable. During the multimodal learning stage, each data modality passes through its pre-trained multiscale ViT backbone to extract the most effective feature maps, which are then utilized for multimodal fusion. We selected three input modalities (RGB, NDVI, and NIR) for multimodal RTS mapping to achieve a robust feature representation. Specifically, the optical image comprises three channels (RGB), while NDVI and NIR each have one channel. To align the input data with the processing requirements of the multi-scale ViT, each modality was rescaled and normalized. Since NDVI and NIR only have one channel, they were duplicated to create three-channel inputs, ensuring compatibility with the multi-scale ViT architecture. Pre-training singular-modality data independently can fully analyze and extract important image features within each modality, avoiding cross-modal disturbances [44]. It can also considerably alleviate memory demands, as will be demonstrated in our following experiments. This approach also sets the stage for subsequent training with flexible combinations of modalities without the need to start the entire process from scratch. Furthermore, it ensures that RTS mapping results, derived from different multimodal combinations, can be compared fairly under a consistent configuration.



(a) RGB (b) NDVI (c) NIR

Figure 3: Example multimodal training data and labels (in red) for RTS segmentation.

4.3 Residual cross-modality attention fusion

Once we obtain a pre-trained model for each modality, we integrate their backbones for multimodal RTS mapping. Each multi-scale ViT backbone processes its designated modality input to generate multi-scale feature maps, denoted as $F_m \in R^{CHW}$, where the feature maps F at each scale for each modality $m \in RGB, NDVI, NIR$ has dimensions HW with C channels. When incorporating N modalities, where $N \in 1, 2, 3$, the combined feature map at each scale has a total of NC channels. Since Cascade Mask R-CNN is originally designed for single-modality input with C channels, it is necessary to adjust the model to accommodate the increased number of channels for multimodal input. Previous studies [36, 44] have adopted a "single convolutional layer" to reduce the number of channels. Feature maps from multiple modalities are concatenated together along the channel dimension and processed through one convolutional layer to create fused feature maps. However, this method has a limited capacity for feature representation, as it could disrupt the learned knowledge of each modality. Meanwhile, the fused feature maps, without the information of individual data modalities, can not fully unleash their potential for leveraging multimodal data [14]. As a result, feature-level fusion using a single convolutional layer could lead to a substantial loss of valuable original information. An efficient fusion strategy aims to harness information from all available modalities, enhancing the performance of a singular modality while minimizing information loss and cross-modal interference. Given the large size of the model, which includes at least one backbone branch per modality, the fusion module must be efficient and lightweight.

To meet these requirements, we developed a feature-level fusion strategy called the "residual cross-modality attention fusion strategy" (refer to Figure 2 for its visual representation). This strategy not only enables the model to learn new features from multiple modalities through the cross-attention operation but also incorporates each modality's original feature information into the new fused feature maps through the residual connection fusion strategy. As a result, more critical information is preserved and enriched with our proposed fusion strategy. During the pre-training stage, we observed that the model using the RGB modality achieved higher prediction accuracy than the other two modalities in a unimodal model. Therefore, our fusion strategy was tailored to enhance the primary RGB modality with either NDVI or NIR as auxiliary resources. This results in two separate fusion processes: one combining NDVI with RGB and the other combining NIR with RGB. In this approach, each supplementary modality (NDVI or NIR) is fused independently without mutual interference. It also provides flexibility for any modality combination that includes RGB without requiring changes to the architecture of the fusion module. When fusing any two modalities, we propose using a cross-attention operation instead of the conventional convolutional fusion. Unlike a self-attention operation, which derives important parameters Q (query), K (key), and V (value) from a single modality, the cross-attention module uses RGB feature patches as the query and feature patches from the other modality (NDVI or NIR) as the key and value. This setup enables RGB to extract complementary information from auxiliary modalities, reinforcing visible-spectrum features in a controlled manner. The outputs of these cross-modality interactions are further combined with residual components to form a composite representation. The design of the residual fusion module is inspired by the seminal convolutional neural network, ResNet [14]. In this network, a residual connection module connects the input of a convolutional block directly to its output, effectively bypassing intermediate layers. This mechanism enables the integration of important information across layers, boosting model performance and overcoming the vanishing gradient problem commonly observed in deep networks. The concept of residual connections has also been adopted in subsequent works, such as DenseNet [16] and ResNeXt [48], further demonstrating its effectiveness in improving deep learning models.

Specifically, when fusing all three modalities in which RGB modality is required, the feature maps of one modality (e.g., NDVI), represented as F_{mod1}, and RGB (F_{rgb}) are fused through a cross-attention layer, $CrossAttn_1$. The process enables the RGB features to dynamically attend to relevant information in the NDVI features, resulting in a fused feature representation F_1'. Similarly, the feature maps of another modality (e.g., NIR), F_{mod1}, and RGB, F_{rgb}, are fused using another cross-attention fusion layer, $CrossAttn_2$, yielding the fused feature F_2', as shown in the following equations:

$$F_1' = CrossAttn_1(F_{rgb}, F_{mod1}) \tag{1}$$

$$F_2' = CrossAttn_2(F_{rgb}, F_{mod2}) \tag{2}$$

The original feature maps for RGB (F_{rgb}) and another two modalities (F_{mod1} and F_{mod2}) are then fused with the previously generated features (F_1' and F_2') through an element-wise addition operation. Finally, a projection layer, $Conv$, is further applied to the result, as shown in the equation:

$$F^{"} = Conv(F_{rgb} \oplus F_{mod1} \oplus F_{mod2} \oplus F_1' \oplus F_2') \tag{3}$$

Where $\oplus$ denotes element-wise addition, and $Conv$ is configured with both input and output channels set to C, the default number of channels in Cascade Mask R-CNN. The resulting feature map, $F^{"}$, serves as the input for subsequent

classification and regression tasks. When the number of input modalities differs from three, the fusion process adapts. Each modality interacts directly with the RGB modality through cross-attention mechanisms. In the case of two modalities, the RGB modality interacts with the other modality via a single cross-attention layer, generating a fused representation. If there are more than three modalities, each additional modality, beyond the RGB, sequentially interacts with the RGB features using its own cross-attention layer. This approach maintains the integrity and dimensions of the feature maps without needing extra layers for specific channel adjustments. Finally, all generated and original feature maps are combined using element-wise addition. A convolutional layer is then applied to this combined result, ensuring the integration of all fused and original features. After the feature fusion, the resulting feature maps are sent to the decoder for instance segmentation (see Figure 2). Specifically, the region proposal network (RPN) uses feature maps to generate object proposals indicating objects' locations [15]. These proposals, along with the feature maps, are then sent to its multi-stage heads, where each head refines results from the previous stage to enrich feature information. In the end, these refined features are employed by the model's segmentation head to generate precise pixel-level masks.

5 Experiements

We evaluated our proposed method using the RTS dataset derived from our study area. The entire RTS dataset comprises 717 training samples and 138 testing images. Each data sample is stored in a GeoTIFF file, containing RGB, NDVI, and NIR data, along with binary masks serving as ground truth. To convert these binary masks into instance-level ground truth, we used OpenCV's "findContours" function, which identifies the contours of distinct objects in the mask. We implemented the model based on the Detectron2 library developed by Facebook AI Research [46]. Detectron2 provides state-of-the-art detection and segmentation algorithms and offers a complete workflow for both training and testing, suitable for research and production applications. For our model, we integrated the proposed multimodal fusion strategy into Cascade Mask R-CNN [2] for instance segmentation. MViTv2 [29] was used as its feature extraction backbone.

Training was conducted on a GPU server with four Nvidia RTX A5000 GPUs, each equipped with 24 GB of memory. Each experiment ran for 36 epochs (complete cycles through the dataset) using the AdamW optimizer [31], selected for its adaptability and regularization benefits. The loss functions were the same as those used in Mask R-CNN, including classification loss (for accurate object labeling), bounding box regression loss (for precise object localization), and mask loss (for detailed per-pixel segmentation). We adjusted anchor sizes to match the typical dimensions of retrogressive thaw slump (RTS) features, allowing the model to better capture objects of interest. In addition, the learning rate was scaled linearly [10] to maintain stable performance as training batch size was dynamically adjusted. For the data split, the same approach was used as described in [52], which involved sampling from each study area to create separate training and testing sets to ensure balanced representation across regions.

The metric employed for evaluating instance segmentation is Average Precision (AP) at an IoU (Intersection over Union) threshold of 0.5, known as AP50. This metric is widely used in instance segmentation tasks to assess a model's ability to accurately detect and localize objects. AP is calculated as the area under the Precision-Recall (P-R) curve, which plots the model's precision and recall at various confidence thresholds. IoU measures the overlap between the predicted mask and the ground truth by dividing the area of intersection by the combined area of both. Specifically, AP50 refers to the Average Precision when the IoU threshold is 0.5. This means a predicted object is considered correct if its IoU with a ground-truth object is greater than 0.5. To compute AP50, precision and recall are evaluated for all predictions, ranked by their confidence scores. Precision is then interpolated to ensure it does not decrease as recall increases, and the area under this interpolated curve is calculated to obtain the AP value. AP50 provides a concise measure of the model's performance at a relatively lenient IoU threshold.

Three groups of experiments were conducted to assess the efficiency of multimodal learning in RTS mapping. To examine the effectiveness of multi-scale ViT enhanced instance segmentation, we compared mapping results across three models: Mask R-CNN with ResNet50-FPN, Mask R-CNN with multi-scale ViT, and Cascade Mask R-CNN with multi-scale ViT. Each model was tested using a single modality input (RGB, NDVI, or NIR). We also assessed models with pre-trained backbones, considering their parameter size and memory consumption. By comparing different modality combinations with various fusion strategies, we not only determined the optimal modality combination that maximized the mapping precision but also indicated that our proposed fusion approach yields the most promising results.

5.1 Model Comparison

As the two popular feature extraction backbones, transformer and CNNs have their own advantages and disadvantages. For example, transformer models excel with large datasets while CNN tends to perform well on smaller datasets [6]. When it comes to transferability, transformers often generalize better across different tasks, even when the downstream data is only weakly related to the data used for pretraining [54, 34]. However, CNNs are generally more

Table 1: Comparison of AP50 (%) on RGB, NDVI, and NIR modality among various model configurations

Model No.	Model Name	Backbone	Modality		
			RGB	NDVI	NIR
1	Mask R-CNN	ResNet50-FPN	31.29	12.78	12.30
2	Mask R-CNN	Multi-scale ViT	37.94	13.57	15.57
3	Cascade Mask R-CNN	Multi-scale ViT	39.73	15.30	18.06

efficient, making them a practical choice for real-time applications or when lightweight backbones are needed [9]. To evaluate the performance of different instance segmentation networks paired with CNN-based (e.g., ResNet50-FPN) and transformer-based backbones (e.g., ViT and multi-scale ViT) for RTS mapping, we tested three models applied on the RGB, NDVI, and NIR modalities respectively. The comparative models included Mask R-CNN with ResNet50-FPN (Model 1), Mask R-CNN with multi-scale ViT (Model 2), and Cascade Mask R-CNN with multi-scale ViT (Model 3). ResNet50-FPN serves as a standard CNN-based benchmark because of its well-known performance on various tasks, making it a useful point of comparison for other models or architectures. It has 50 layers and uses residual connections to capture representative features. Meanwhile, Feature Pyramid Networks (FPN) combined with ResNet50 further creates multi-level and hierarchical representation of features at various scales. ViT and multi-scale ViT are transformer-based backbones and they are integrated into the instance segmentation pipelines Mask R-CNN and Cascade Mask R-CNN, respectively. Table 1 shows the performance statistics of these models.

The results indicate that the RGB modality yields the highest mapping accuracy compared to NDVI and NIR across all three model configurations. When using RGB as the input, the Mask R-CNN with a multi-scale ViT backbone (Model 2) achieves a prediction accuracy of 37.94%, representing an improvement of 6.65% over the Mask R-CNN with a ResNet50-FPN backbone (Model 1), which has an AP50 of 31.29%. This performance advantage is also observed when the other two modalities are used as input, with a performance increase from 12.78% to 13.57% for NDVI and from 12.30% to 15.57% for NIR, respectively. Model 2's better performance compared to Model 1 is attributed to the capabilities of the multi-scale ViT in extracting important feature representations by capturing data dependencies not only across multiple scales but also over long ranges. A comparison between different instance segmentation networks, Mask R-CNN and Cascade Mask R-CNN, was conducted. Both networks used the multi-scale ViT as the backbone, as it outperforms CNN-based backbones for this specific task. For all data modalities, Cascade Mask R-CNN (Model 3) achieved higher performance than both Models 1 and 2, demonstrating its superiority in model architecture by adopting multi-stage mask refinement to improve segmentation results. Based on these experimental results, our study builds upon and extends Model 3 to enable multimodal learning using the proposed residual cross-modality attention fusion strategy.

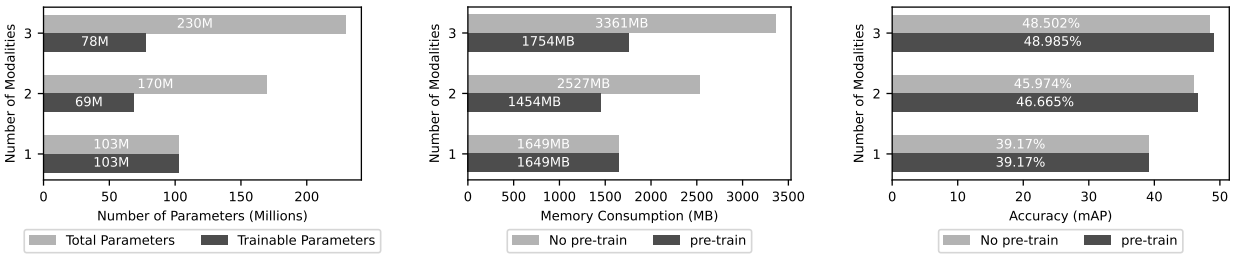
5.2 Model performance with unimodal pretraining and multimodal fine tuning

In multimodal training, leveraging the weights of pre-trained backbones without further training has proven to be efficient, as these weights have already achieved the highest AP50 for their respective modalities. Table 2 presents an analysis of the model's parameters, memory consumption, and accuracy, comparing models with and without pre-trained backbones, with the differences further illustrated in Figure 4. For the memory consumption, it is estimated based on the peak memory usage observed during training. This includes the memory required for loading one data sample, loading the model, performing the forward pass, and executing the backward pass. Other potential sources of memory usage, such as optimizer states (e.g., moment estimates in Adam), temporary buffers, CUDA memory caching, and data augmentation, are excluded from this estimation. This consistent baseline provides an accurate comparison of the computational demands for models with and without pre-trained backbones.

For instance, segmentation models trained on a single modality, such as RGB, NDVI, or NIR, do not require pre-training. These models consist of approximately 103 million (M) parameters, all of which are trainable during the training process. Consequently, the total number of trainable parameters remains 103M, with a memory usage of approximately 1649MB. In contrast, when two modalities (e.g., RGB and NDVI, or RGB and NIR) are utilized, the inclusion of two pre-trained backbones increases the total parameter count to 170M (Table 2). However, since these backbones are set to be non-trainable, the number of trainable parameters decreases significantly to 69M. This pre-training strategy also reduces memory consumption, requiring only 1454MB compared to 2527MB if all parameters were trainable. While these differences may seem modest, the actual memory usage during training can increase significantly—up to fivefold—when data augmentation is applied. This is due to the quadratic relationship between transformer memory usage and patch sequence length O(n2), in addition to other memory overheads. In terms of performance, the pre-

Table 2: Model performance and resource usage for different modalities, detailing parameters, memory, and accuracy with and without pre-training.

Number of Modalities	Number of Total Parameters (Millions)	Number of Trainable Parameters (Millions)	Memory (MB)		Accuracy (AP %)	
			No Pre-training	With Pre-training	No Pre-training	With Pre-training
1	103	103	1649		39.17	
2	170	69	2527	1454	45.97	46.67
3	230	78	3361	1754	48.50	48.99



(a) Model Parameters (b) Memory Consumption (c) Accuracy

Figure 4: Statistics on model parameters, memory consumption and accuracy (measured by average precision).

training strategy proved advantageous, achieving a slightly higher AP score of 46.67% compared to 45.97% for the non-pre-trained model.

Similarly, integrating all three data modalities (RGB, NDVI, and NIR) leads to a linear increase in the total number of parameters to 230M, but only 78M are trainable when leveraging pre-trained backbones. This configuration offers a substantial reduction in memory requirements, consuming just 1754MB compared to 3361MB when all parameters are trainable. The memory savings become even more critical in practical training scenarios where factors like data augmentation and increased patch sequence lengths can significantly inflate memory usage. From a performance perspective, incorporating pre-trained backbones once again proves beneficial, achieving an accuracy (AP) of 48.99%, slightly outperforming the 48.50% achieved without pre-training. These results highlight the scalability and efficiency of the pretraining approach as the number of modalities increases. They also demonstrate that the proposed training strategy maintains comparable or even slightly better performance than the direct training approach, reinforcing its practical applicability.

5.3 Effectiveness of multimodal fusion strategies for RTS mapping

In this experiment, we aimed to evaluate the accuracy of RTS mapping utilizing different multi-modalities and fusion methods. Specifically, we compared the efficacy of our proposed residual cross-modality attention fusion strategy with data-level fusion [52], feature-level convolutional fusion [36, 44], stacked-modality attention fusion [50], and cross-modality attention fusion [33, 51, 32, 20].

1. data-level fusion: this is the most commonly adopted approach for incorporating multimodality data, especially in RTS mapping applications [52]. In this approach, different modality data are concatenated along the channel dimension. While simple and easy to implement, this approach requires careful alignment and preprocessing input modalities to ensure consistent spatial resolution. Figure 5a illustrates the architecture of data-level fusion.
2. feature-level convolutional fusion: this strategy introduces a convolutional layer that takes the feature maps extracted from each data modality as input, and applies a convolution operation to fuse them into a unified representation. By stacking the feature maps from individual backbones, this method merges complementary information from each modality. Feature-level fusion has been used to detect landform features, such as valleys and mountains [44]. Figure 5b shows an example architecture design for feature-level convolutional fusion.
3. stacked-modality attention fusion: this is a feature-level fusion strategy. The input of the fusion module includes feature maps generated from each modality, and then they are stacked to become a comprehensive

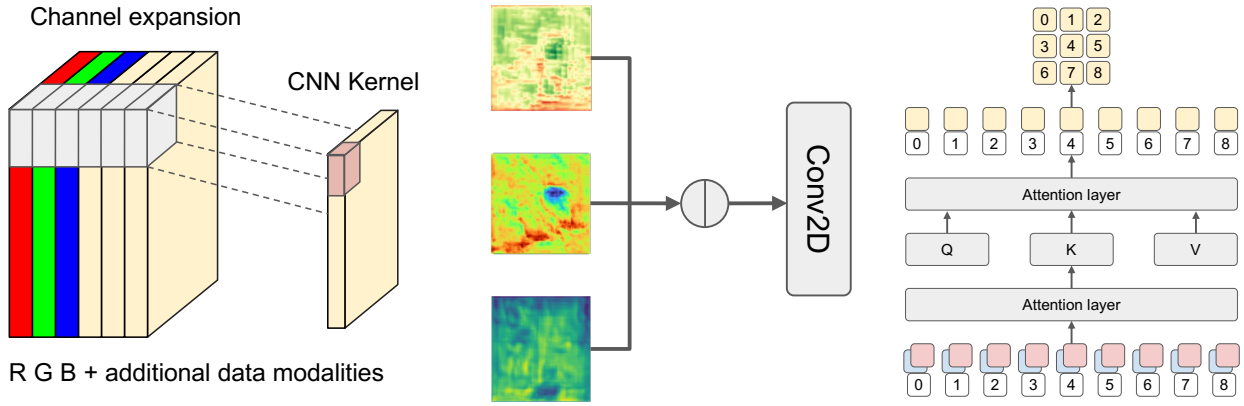


Figure 5: Design and implementation of different fusion strategies. (a) Data-level fusion through channel expansion. (b) Feature-level convolutional fusion. Conv2D means a 2D convolutional layer. (c) Stacked-modality attention fusion. Q, K, and V represent the query, key, and value embeddings for attention calculation.

feature representation. After that, an attention module is applied to extract feature maps by dynamically analyzing relevant features across the stacked modalities [43, 50]. Figure 5c illustrates an example design for stacked-modality attention fusion.

4. cross-modality attention fusion: in cross-modal attention fusion, instead of applying the attention module on a single modality, query embeddings from one modality (RGB) will operate with keys and values from other modalities (e.g., NDVI and NIR), enabling conditioned interactions across modalities. This allows the model to focus on relevant features in the context of the RGB data, leveraging the complementary information from other modalities. Unlike stacked-modality attention, where the attention module processes a combined stack of all feature maps, cross-modality attention specifically enhances the understanding of one modality by dynamically attending to the complementary features provided by the others. Cross-modality attention has been effectively used in hyperspectral and lidar classification, as demonstrated by [33] and [51]. The subfigure labeled with "Fusion module" in Figure 2 shows our proposed design and implementation for cross-modality attention fusion.

Table 3 provides the model prediction accuracy (measured by AP50) using different fusion strategies and modality combinations. The RGB modality alone is used as a baseline to assess the efficiency of multimodal combinations, given that no fusion strategy is employed in this scenario. The AP50 value of 39.73% yielded from the singular RGB modality can be regarded as our baseline against which other results can be compared. Employing a data-level fusion strategy with the RGB and NDVI modalities achieves the AP50 of 42.74%, which surpasses the result from the singular RGB modality. When the NIR modality is added to this combination, the accuracy is further improved to 43.87%. This result demonstrates the added value of multimodal data in providing complementary information to enhance the model's predictive capability for delineating RTS.

Table 3: AP50 (%) results for multimodal combinations with various fusion strategies.

Fusion Type	Fusion Strategy	Input Modalities		
		RGB	RGB + NDVI	RGB + NDVI + NIR
Data level	Channel expansion	39.73	42.74	43.87
Feature level	Convolutional fusion		44.51	45.45
	Stacked-modality attention fusion		44.53	45.79
	Cross-modality attention fusion		44.71	46.72
	Residual cross-modality attention fusion (our proposed)		46.67	48.99

Furthermore, all of the comparative feature-level fusion methods outperform data-level fusion across all multimodal combinations. Overall, our proposed residual cross-modality attention fusion approach achieves the highest prediction accuracy of 46.67% when two modalities (RGB and NDVI) are combined and 48.99% when three modalities (RGB, NDVI, and NIR) are used in combination (Table 3). The advantage of our proposed strategy is attributed to the introduction of both the residual connection module and the cross-modality attention module. The residual connection combines original feature maps from different modalities with fused multimodal feature maps, enabling the model to leverage information from both sources for improved performance. Meanwhile, the cross-modality attention module allows the direct integration of complementary information from other modalities (NDVI and NIR) into the RGB feature space. By integrating features in this way, the model can more effectively capture richer cross-modal dependencies, resulting in superior performance in multimodal learning.

The second-best performing strategy is cross-modality attention fusion (without the residual connection). The convolutional fusion and stacked-modality attention fusion strategies achieve similar levels of performance for both modality combinations, and both outperform data-level fusion. This is because when each modality is processed independently to generate feature maps, the inherent characteristics and unique information of each modality are preserved without cross-modal interference. In contrast, channel expansion in data-level fusion blends features from various modality distributions, potentially introducing noise into the model and degrading performance.

To further evaluate the effectiveness of our proposed fusion strategy, we conducted an ablation study by replacing the cross-modality attention strategy in the residual connection fusion with convolutional fusion and stacked-modality attention fusion. Table 4 shows the experimental results using different modality inputs. By cross-comparing with the results in Table 3, we observe that residual connection fusion, even with the commonly used convolutional fusion to combine multi-modality information, results in better predictive performance (AP50 of 46.21% when combining RGB and NDVI, and 47.62% when combining RGB with NDVI and NIR) than using attention-based fusion strategies alone.

When stacked-modality attention fusion is used without the residual connection module, the model achieves a prediction accuracy of 44.53% when using RGB and NDVI as input, and 45.79% when using all three modalities (RGB, NDVI, and NIR) as input (Table 3). The cross-modality attention fusion used in the same scenario achieves a prediction accuracy of 44.71% when using RGB and NDVI as input, and 46.72% when using all three modalities (RGB, NDVI, and NIR) as input (Table 3). Both approaches underperformed compared to residual convolutional fusion (Table 4). This finding highlights the effectiveness of residual connection fusion in multimodal learning. Meanwhile, after integrating stacked-modality attention fusion and cross-modality attention fusion, the model performance is further enhanced compared to residual connection fusion, reflecting the effectiveness of attention fusion strategies in multimodal information extraction (Table 4). In particular, our proposed residual cross-modality attention fusion achieves the best overall predictive performance at 48.99% when three modality data are used, demonstrating its capability to fully integrate complementary information from multiple modality data.

5.4 Visualization of RTS mapping results

Figure 6 presents examples of RTS feature predictions. The ground truths are marked in red, and predicted instances are outlined in blue. The first column (a) displays the ground truth instances, while subsequent columns present predictions from one data-level and two feature-level fusion approaches. These approaches utilize all data modalities including RGB, NDVI and NIR. Overall, the results demonstrate that the proposed residual cross-modality fusion outperforms the other two methods. It achieves both recognizing all instances, even when multiple are present, and more precise boundary detection for each instance.

In the first example, RTS01, all three methods successfully detect the two instances. However, both data level fusion (b) and convolutional fusion (c) mistakenly predict a larger extent for one of the RTS instances. In contrast, the residual cross-modality fusion (d) provides a more precise detection for both instances, closely aligning with the ground truth boundaries. In the second example, RTS02, the ground truth displays multiple instances outlined in red. Both data level fusion and convolutional fusion struggle, detecting only one instance. In contrast, the residual cross-modality fusion

Table 4: AP50 (%) results for ablation study

Fusion Strategy	Input Modalities		
	RGB	RGB + NDVI	RGB + NDVI + NIR
Residual convolutional fusion	39.73	46.21	47.62
Residual stacked-modality attention fusion		47.00	47.69
Residual cross-modality attention Fusion		46.67	48.99

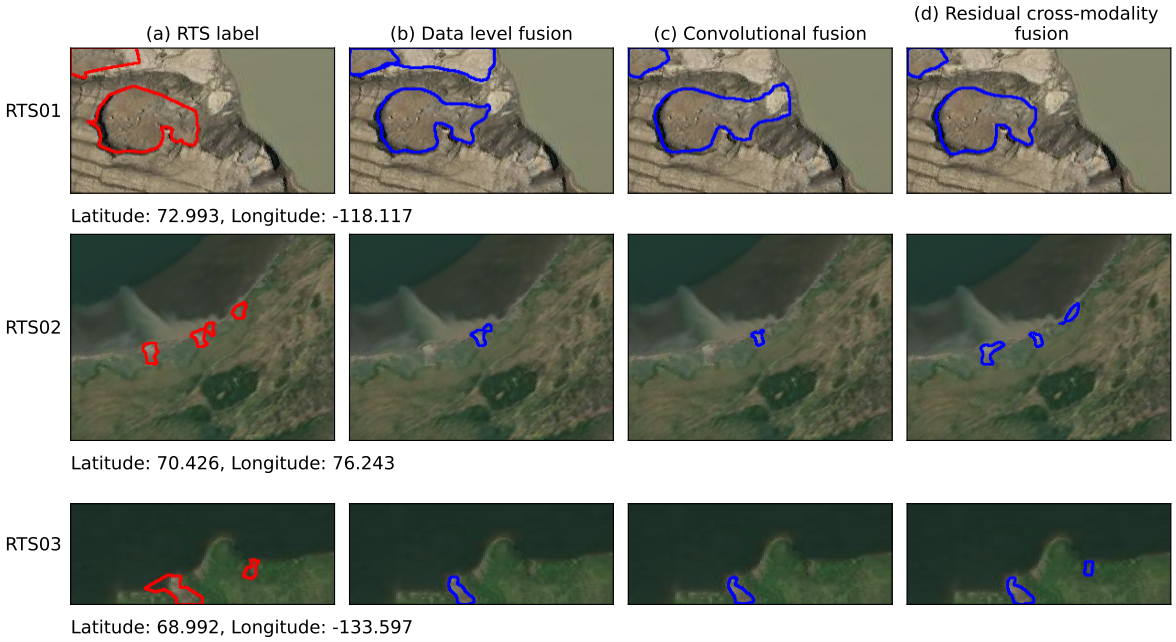


Figure 6: Visualization of different model prediction results using combinations of all data modalities.

effectively detects all instances. Although its boundary accuracy still needs improvement, it demonstrates a stronger ability to recognize general RTS features. Similarly, in the last example, RTS03, the residual cross-modality fusion identifies all instances while the other two methods only capture one instance.

In conclusion, from the three examples, all three methods are capable of detecting RTS features when they are larger. However, when the features are smaller, only our approach, the residual cross-modality fusion, consistently succeeds in detection.

Figure 7 illustrates the impact of various data modality combinations on the detection of RTS features using our proposed method, cross-modality attention fusion. The first column (a) shows the ground truth RTS labels in red. The subsequent columns depict detection results: column (b) uses RGB data, column (c) combines RGB and NDVI, and column (d) integrates RGB, NDVI and NIR. Blue outlines indicate detected features. The progression across columns demonstrates how additional data modalities enhance the model's ability to detect RTS features.

In the first example, RTS04, using only RGB data provides limited precision. Adding NDVI enhances accuracy and the inclusion of NIR achieves the most precise detection, identifying all three RTS features as the ground truth. Comparing this to Figure 6, RTS02, which uses the same sample, illustrates that adding more data modalities and effectively fusing them are equally important for achieving optimal results.

In this RTS06 example, similar to RTS04, the aim is to identify all instances accurately. The results for RTS06 are consistent with RTS04, where integrating more data modalities leads to improved detection accuracy.

In the RTS05 example, the initial use of RTS data results in duplicate detections, where multiple outlines overlap the same features. It also identifies additional instances compared to the ground truth. Adding NDVI reduces the extra detection but still shows some duplications. The integration of all modalities further refines the detection, effectively eliminating duplicates, extras and aligning closely with the ground truth.

Overall, the figure demonstrates that incorporating more data modalities improves the RTS detection. As additional modalities like NDVI and NIR are included, the model shows enhanced capability in identifying all instances with greater accuracy. This results in more precise boundary delineations and reduces the occurrence of duplicate instances, highlighting the effectiveness of the cross-modality attention fusion method.

5.5 Model generalizability test

To verify the effectiveness of our proposed multimodal feature fusion strategy, we extended the evaluation on another multimodal landform benchmark dataset, the GeoImageNet [28]. GeoImageNet is a natural feature dataset that contains

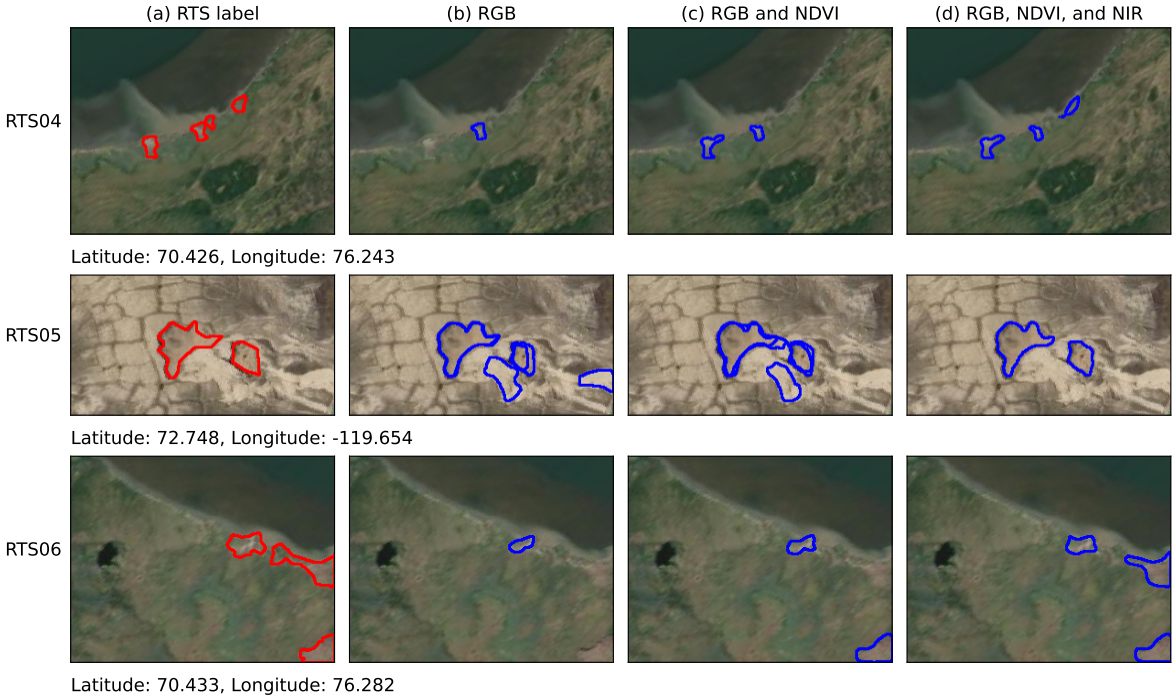
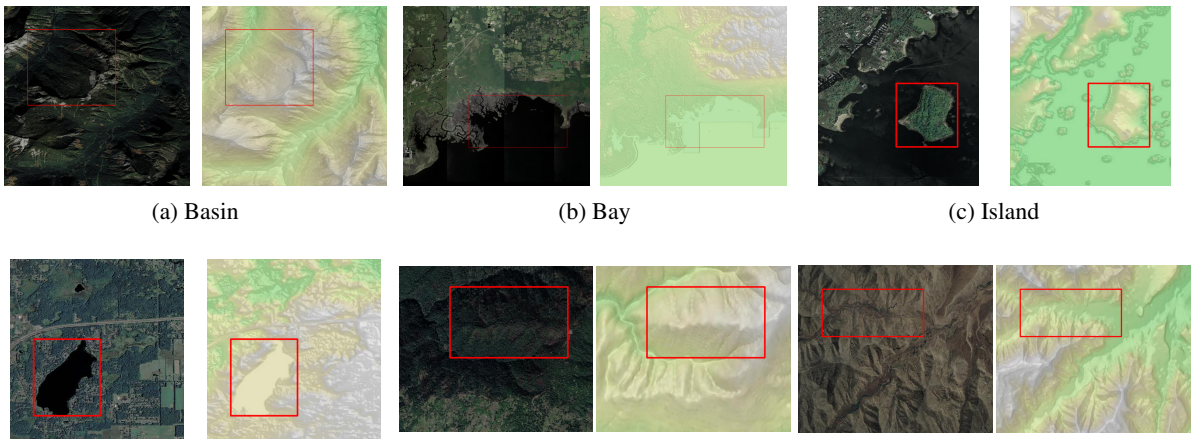


Figure 7: Visualization of model prediction results using different data modality combinations.

six types of terrain features, including basin, bay, island, lake, ridge, and valley, in both hilly and flat terrains. The training images contain data from two modalities, 1) optical images with 3 RGB bands from 1-meter resolution National Agriculture Imagery Program (NAIP) aerial imagery, and 2) an enhanced DEM data derived from 1/3 arc-second (approximately 10-meter) resolution USGS 3D Elevation Program (3DEP) data. The original DEM data with numerical values are enhanced by compositing multiple rendered DEM derivatives (including color relief, slope, and hill shade) to better align with deep learning paradigms. The features are randomly sampled from the GNIS database, covering 44 states of continental US and Hawaii. There are a total of 876 images labeled by experts through visual inspection, referencing both the USGS Historical Topographic Map Collection (HTMC) data set and NAIP imagery.

Similar to thaw slump dataset, GeoImageNet is also an AI-ready dataset for landform mapping. Figure 8 shows some example training images with its annotations. Different from the thaw slump dataset, the GeoImageNet data is labeled



(a) Basin (b) Bay (c) Island

(d) Lake (e) Ridge (f) Valley

Figure 8: Sample training images with their annotations from GeoImageNet.

Table 5: AP50 (%) results for ablation study

Fusion Type	Fusion Strategy	Input Modalities	
		RGB	RGB + DEM
Data level	Channel Expansion	73.39	82.88
Feature level	Convolutional Fusion		85.28
	Residual Convolutional Fusion		85.78
	Residual Stacked-Attention Fusion		86.50
	Residual Cross-Attention Fusion		86.68

to support object detection tasks, which include object localization by predicting a bounding box and classification of object class. This variation allows us to evaluate our model's performance and robustness not only among different datasets but also different image analysis tasks. Fortunately, the instance segmentation framework we adopted can easily be adapted to support object detection, by using a different head (decoder). So almost the entire image analysis pipeline, including the feature fusion module, illustrated in Figure 2 can be reused. The only difference is the removal of the mask head from the Cascade Mask R-CNN architecture. The object detection head is retained to directly predict bounding boxes and class labels.

Table 5 shows the model prediction performance using different fusion strategies and data modalities. It can be observed that fusing RGB data with an additional modality—the enriched DEM—substantially increases the model's predictive accuracy. It is evident that feature-level fusion strategies are more effective than data-level fusion achieved through channel expansion. Specifically, residual convolutional fusion, which combines the original feature maps with those generated after fusion, results in stronger feature representation and outperforms the commonly used convolutional feature fusion method. Further enhancing this approach with an attention fusion module continues to boost performance. Overall, the residual cross-attention fusion achieves the best predictive performance, with a nearly 14% increase (from 73.39% to 86.68%) in AP compared to single-modal learning results and a 3.8% increase (from 82.88% to 86.68%) compared to the data-level fusion strategy. This finding is consistent with experimental results on the retrogressive thaw slump dataset, providing further evidence of the generalizability and robustness of our proposed feature fusion strategy. It also validates the applicability of our method across varied datasets, geographical contexts (e.g., pan-Arctic vs. the contiguous US), and image analysis tasks (e.g., object detection vs. instance segmentation).

6 Conclusion

In this paper, we developed a multimodal approach for mapping retrogressive thaw slumps across the Arctic. Two innovative features were introduced to achieve strong model performance. First, we proposed a residual cross-modality attention fusion strategy that integrates the advantages of (1) a residual connection module, which leverages incremental features important for segmenting thaw slump features by enabling information bypass, and (2) a cross-modality attention layer, which enhances the resultant feature map with complementary information from different modalities through the cross-attention mechanism. Second, to address the high memory consumption introduced by multimodality processing and fusion, we introduced a training strategy that adopts unimodality pre-training in Stage 1 and multimodality fine-tuning with a frozen backbone in Stage 2. This approach enables the model to achieve comparable performance with significantly lower memory demands than regular multimodal training. This training strategy also allows the model to be easily expanded to incorporate additional modalities without the need for re-pretraining, ensuring high scalability for multimodal models and improved computing efficiency in terms of resource consumption. In today's era of large AI models, developing efficient yet powerful models is critical for addressing geospatial challenges, such as permafrost thaw mapping, while preserving AI's environmental friendliness [25]. The superior performance of our proposed model and multimodal fusion strategy is validated not only on the thaw slump datasets but also through its generalizability, as demonstrated on another landform dataset, GeoImageNet.

In the future, we plan to incorporate additional data modalities into our multimodal permafrost thaw framework to further enhance the model's robustness across heterogeneous Arctic landscapes. In particular, we will evaluate the effectiveness of the ArcticDEM available from the Polar Geospatial Center to support our mapping work. DEM data captures elevation changes and provides the vertical profile of terrain information, making it highly useful for mapping retrogressive thaw slumps, which are formed by ground subsidence and often involve the collapse of land from its surroundings. However, because thaw slumps can be small in scale with minor vertical drop, the DEM data must possess high vertical accuracy in addition to the desired spatial resolution to support high-accuracy RTS mapping. In addition to continuing multimodal modeling, we also plan to operationalize thaw slump mapping at a pan-Arctic scale.

Such a high-resolution dataset is critical for understanding permafrost thaw and for developing new AI-based models to better identify its triggering factors and conduct near-term forecasts of abrupt permafrost thaw. This research, which integrates AI with domain science, will further advance scientific inquiry and foster the development of new knowledge to better understand the world's changing environment.

Acknowledgment

This research is supported in part by Google.org's Impact Challenge for Climate Innovation Program and the National Science Foundation under awards 2120943, 2230034, 2230035.

References

[1] K. Bayoudh, R. Knani, F. Hamdaoui, and A. Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38(8):2939–2970, 2022.

[2] Z. Cai and N. Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1483–1498, 2019.

[3] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision–ECCV 2018*, pages 833–851. Springer, 2018.

[5] P. Chylek, C. Folland, J. D. Klett, M. Wang, N. Hengartner, G. Lesins, and M. K. Dubey. Annual mean arctic amplification 1970–2020: observed and simulated by cmip6 climate models. *Geophysical Research Letters*, 49(13):e2022GL099371, 2022.

[6] 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. In *9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

[7] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6804–6815, 2021.

[8] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[9] M. Goldblum, H. Souri, R. Ni, M. Shu, V. Prabhu, G. Somepalli, P. Chattopadhyay, M. Ibrahim, A. Bardes, J. Hoffman, R. Chellappa, A. G. Wilson, and T. Goldstein. Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] P. Goyal, P. Dollár, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *CoRR*, abs/1706.02677, 2017. Retrieved from `http://arxiv.org/abs/1706.02677` (Last accessed on November 11, 2024).

[11] S. Hayes, M. Lim, D. Whalen, P. J. Mann, P. Fraser, R. Penlington, and J. Martin. The role of massive ice and exposed headwall properties on retrogressive thaw slump activity. *Journal of Geophysical Research: Earth Surface*, 127:e2022JF006602, 2022.

[12] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016*, pages 630–645. Springer, 2016.

[15] C.-Y. Hsu, W. Li, and S. Wang. Knowledge-driven geoai: Integrating spatial knowledge into multi-scale deep learning for mars crater detection. *Remote Sensing*, 13(11):2116, 2021.

[16] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[17] L. Huang, T. C. Lantz, R. H. Fraser, K. F. Tiampo, M. J. Willis, and K. Schaefer. Accuracy, efficiency, and transferability of a deep learning model for mapping retrogressive thaw slumps across the canadian arctic. *Remote Sensing*, 14(12):2747, 2022.

[18] L. Huang, J. Luo, Z. Lin, F. Niu, and L. Liu. Using deep learning to map retrogressive thaw slumps in the beiluhe region (tibetan plateau) from cubesat images. *Remote Sensing of Environment*, 237:111534, 2020.

[19] Z. Huang, C. Lv, Y. Xing, and J. Wu. Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. *IEEE Sensors Journal*, 21(10):11781–11790, 2020.

[20] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning*, pages 4651–4664. PMLR, 2021.

[21] M. K. Jones, W. H. Pollard, and B. M. Jones. Rapid initialization of retrogressive thaw slumps in the canadian high arctic and their response to climate and terrain factors. *Environmental Research Letters*, 14(5):055006, 2019.

[22] H. Lantuit and W. H. Pollard. Fifty years of coastal erosion and retrogressive thaw slump activity on herschel island, southern beaufort sea, yukon territory, canada. *Geomorphology*, 95(1-2):84–102, 2008.

[23] H. Lee and W. Li. Improving interpretability of deep active learning for flood inundation mapping through class ambiguity indices using multi-spectral satellite imagery. *Remote Sensing of Environment*, 309:114213, 2024.

[24] W. Li. Geoai: Where machine learning and big data converge in giscience. *Journal of Spatial Information Science*, (20):71–77, 2020.

[25] W. Li, S. Arundel, S. Gao, M. Goodchild, Y. Hu, S. Wang, and A. Zipf. Geoai for science and the science of geoai. *Journal of Spatial Information Science*, (29):1–17, 2024.

[26] W. Li and C.-Y. Hsu. Geoai for large-scale image analysis and machine vision: recent progress of artificial intelligence in geography. *ISPRS International Journal of Geo-Information*, 11(7):385, 2022.

[27] W. Li, C.-Y. Hsu, S. Wang, C. Witharana, and A. Liljedahl. Real-time geoai for high-resolution mapping and segmentation of arctic permafrost features: the case of ice-wedge polygons. In D. D. Lunga and S. D. Newsam, editors, *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, GeoAI 2022*, pages 62–65, Seattle, Washington, 2022.

[28] W. Li, S. Wang, S. T. Arundel, and C.-Y. Hsu. Geoimagenet: a multi-source natural feature benchmark dataset for geoai and supervised machine learning. *GeoInformatica*, 27(3):619–640, 2023.

[29] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4804, 2022.

[30] 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 (ICCV)*, pages 9992–10002, 2021.

[31] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[32] S. Lu, M. Liu, L. Yin, Z. Yin, X. Liu, and W. Zheng. The multi-modal fusion in visual question answering: a review of attention mechanisms. *PeerJ Computer Science*, 9:e1400, 2023.

[33] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 416–425, 2020.

[34] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[35] I. Nitze, K. Heidler, S. Barth, and G. Grosse. Developing and testing a deep learning approach for mapping retrogressive thaw slumps. *Remote Sensing*, 13(21):4294, 2021.

[36] T. Ophoff, K. Van Beeck, and T. Goedemé. Exploring rgb+ depth fusion for real-time object detection. *Sensors*, 19(4):866, 2019.

[37] J. L. Ramage, A. M. Irrgang, U. Herzschuh, A. Morgenstern, N. Couture, and H. Lantuit. Terrain controls on the occurrence of coastal retrogressive thaw slumps along the yukon coast, canada. *Journal of Geophysical Research: Earth Surface*, 122(9):1619–1634, 2017.

[38] 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*, pages 234–241. Springer, 2015.

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015. Retrieved from `http://arxiv.org/abs/1409.1556`.

[40] K. W. Turner, M. D. Pearce, and D. D. Hughes. Detailed characterization and monitoring of a retrogressive thaw slump from remotely piloted aircraft systems and identifying associated influence on carbon and nitrogen export. *Remote Sensing*, 13(2):171, 2021.

[41] M. Udawalpola, C. Witharana, A. Hasan, A. Liljedahl, M. Ward Jones, and B. Jones. Automated recognition of permafrost disturbances using high-spatial resolution satellite imagery and deep learning models. In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 46, pages 203–208, 2022.

[42] P. Ulmas and I. Liiv. Segmentation of satellite imagery using u-net models for land cover classification. *arXiv preprint arXiv:2003.02899*, 2020.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] S. Wang and W. Li. Geoai in terrain analysis: Enabling multi-source deep learning and data fusion for natural feature detection. *Computers, Environment and Urban Systems*, 90:101715, 2021.

[45] C. Witharana, M. R. Udawalpola, A. K. Liljedahl, M. K. W. Jones, B. M. Jones, A. Hasan, D. Joshi, and E. Manos. Automated detection of retrogressive thaw slumps in the high arctic using high-resolution satellite imagery. *Remote Sensing*, 14(17):4132, 2022.

[46] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. Last accessed on November 10, 2024.

[47] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3974–3983, 2018.

[48] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492–1500, 2017.

[49] J. Xu, J. Yang, X. Xiong, H. Li, J. Huang, K. C. Ting, Y. Ying, and T. Lin. Towards interpreting multi-temporal deep learning models in crop mapping. *Remote Sensing of Environment*, 264:112599, 2021.

[50] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[51] B. Yang, X. Wang, Y. Xing, C. Cheng, W. Jiang, and Q. Feng. Modality fusion vision transformer for hyperspectral and lidar data collaborative classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:17052–17065, 2024.

[52] Y. Yang, B. M. Rogers, G. Fiske, J. Watts, S. Potter, T. Windholz, A. Mullen, I. Nitze, and S. M. Natali. Mapping retrogressive thaw slumps using deep neural networks. *Remote Sensing of Environment*, 288:113495, 2023.

[53] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021.

[54] H. Y. Zhou, C. Lu, S. Yang, and Y. Yu. Convnets vs. transformers: Whose visual representations are more transferable? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2230–2238, 2021.

[55] S. Zolkos, G. Fiske, T. Windholz, G. Duran, Z. Yang, V. Olenchenko, A. Faguet, and S. M. Natali. Detecting and mapping gas emission craters on the yamal and gydan peninsulas, western siberia. *Geosciences*, 11(1):21, 2021.