

**Key Points:**

- Semantic segmentation of Retrogressive Thaw Slumps (RTSs) is limited by the generalization ability
- Our proposed Unsupervised Domain Adaptation method achieved 14.32%–24.17% increase of F1 in RTS mapping compared with classic methods
- The RTSs in Beiluhe are rapidly expanding and their distribution correlates with the topographic and geological environments

**Correspondence to:**

X. Hu,  
hu.xie@pku.edu.cn

**Citation:**

Lin, Y., Hu, X., Lu, H., Niu, F., Liu, G., Huang, L., et al. (2025). Multi-annual inventorying of retrogressive thaw slumps using domain adaptation. *Journal of Geophysical Research: Machine Learning and Computation*, 2, e2024JH000370.  
<https://doi.org/10.1029/2024JH000370>

Received 26 JUL 2024  
Accepted 2 DEC 2024

## Multi-Annual Inventorying of Retrogressive Thaw Slumps Using Domain Adaptation

Yiling Lin<sup>1</sup> , Xie Hu<sup>1</sup> , Haoyu Lu<sup>2</sup>, Fujun Niu<sup>3</sup> , Gengnian Liu<sup>1</sup>, Lingcao Huang<sup>4</sup>, Shanghang Zhang<sup>5</sup>, Jifu Liu<sup>6</sup>, and Yunhuai Liu<sup>5</sup>

<sup>1</sup>College of Urban and Environmental Sciences, Peking University, Beijing, China, <sup>2</sup>School of Artificial Intelligence, Beijing Normal University, Beijing, China, <sup>3</sup>Civil Engineering at the State Key Laboratory of Frozen Soil Engineering, Northwest Institute of Eco-environment and Resources, Chinese Academy of Sciences, Gansu, China, <sup>4</sup>Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Hong Kong, China, <sup>5</sup>The School of Computer Science, Peking University, Beijing, China, <sup>6</sup>Faculty of Geographical Science, Beijing Normal University, Beijing, China

**Abstract** Retrogressive Thaw Slumps (RTSs), a form of thermokarst hazards, pose risks to hydrological and ecological environments and the safety of the Qinghai-Tibet Engineering Corridor. We still lack the knowledge about the geographic locations of RTSs and their dynamically changing spatial margins. However, visual interpretation is labor-intensive while the present-day deep learning methods become ineffective when the model trained in one year is directly transferred to another. To enhance the model's generalization ability, here we implemented and compared three domain adaptation methods, that is, the classic supervised fine-tuning method and two proposed unsupervised methods: Image StyleTransfer Domain Adaptation (ISTDA) and the Tversky Adversarial Domain Adaptation (TADA) network. In our proposed ISTDA, we uniformed the contextual information of multi-temporal images by Cycle Generative Adversarial Network (CycleGAN). We introduced the Tversky loss and the automatic adjustment of weights for multiple loss functions to suppress false positives and to improve the generalization of TADA. We tested three methods' performance in Beiluhe region over the Qinghai-Tibet Plateau using PlanetScope optical images during 2019–2022. The three domain adaptation methods are successful in generating regional, multi-annual RTS inventories. Remarkably, TADA sustains good performance in complex transfer scenarios without additional label cost, achieving an F1 increase of 14.32%–24.17% compared to classic methods. Our work is the first to apply an unsupervised domain adaptation to automatically map the RTSs on a multi-annual timescale, demonstrating a strong potential of its applicability for monitoring large-scale, multi-temporal evolution of geomorphological features.

**Plain Language Summary** The Retrogressive Thaw Slumps (RTSs) are one of the thermokarst hazards that keeps expanding during warming climate. Knowledge about the geographic locations and the dynamic changes in the exterior boundaries of RTSs is crucial for assessing their impact on the natural environment and for protecting transportation corridors nearby. However, when it comes to the multi-annual inventorying, the present-day RTS mapping algorithms might be ineffective. Our proposed methods, the Image Style Transfer Domain Adaptation and the Tversky Adversarial Domain Adaptation, significantly improve our ability to recover the evolution of RTSs accurately. In this study, the RTS boundaries are delineated on an annual basis from 2019 to 2022 in Beiluhe region in the Qinghai-Tibet Plateau. RTSs mostly occurred on gentle slopes facing to the north and loam soil with moderate to high moisture.

### 1. Introduction

Since the 1980s, the temperature over the Qinghai-Tibet Plateau has been increasing, resulting in various forms of thermokarst landforms, such as Retrogressive Thaw Slumps (RTSs) and thermokarst lakes (Cheng & Wu, 2007; Jorgenson, 2013; Kuang & Jiao, 2016; Li et al., 2019; Walter et al., 2007). RTSs are slope failures initiated from the removal of the protective active layer, water erosion, and anthropogenic disturbance and the consequent exposure of ice-rich permafrost. Besides the downslope mass wasting, the RTSs keep expanding upslope due to a continue exposure of permafrost and underground ice melts (Bernhard et al., 2022; Burn & Lewkowicz, 1990; Ramage et al., 2017; Runge et al., 2022; Zwieback et al., 2018). In recent years, RTSs in the Qinghai-Tibet Plateau are highly active with a significant increase in both their numbers and the affected areas (Luo et al., 2019; Lu et al., 2023; D. Yang, Qiu, et al., 2023). The melting of underground ice during RTSs' development promotes sediments, solute, and water into streams, disturbing the hydrological environment (Kokelj et al., 2021; Rudy

et al., 2017). Meanwhile, the organic carbon stored in the exposed permafrost may be released and notably affects the global climate (Mu et al., 2020; Turetsky et al., 2020). In particular for the Qinghai-Tibet Plateau, RTSs threaten the infrastructure stability (Hjort et al., 2022; Ran et al., 2022).

The substantial risks that RTSs present to ecological systems and infrastructure have made the creation of precise RTS inventories an urgent issue. The high cost and subjective nature of manual visual interpretation, pose challenges in accurately identifying a comprehensive array of RTSs over extensive regions. Recent advancements in artificial intelligence have led to the widespread development of automatic RTS mapping algorithms that utilize deep learning techniques (e.g., Huang et al., 2020; Nitze et al., 2021; Y. Yang, Rogers, et al., 2023). Previous studies have demonstrated the feasibility of deep learning in identifying single-temporal RTSs in parts of the Qinghai-Tibet Plateau (e.g., Xia et al., 2022).

However, the ever-changing characteristics of RTSs render single-temporal inventories inadequate for monitoring their development, exploring the fundamental driving forces, evaluating future environmental conditions, and alleviating engineering impacts. Multi-temporal RTS inventories enable quantitative estimation of RTS occurrence and retreat rates (Huang et al., 2021). Furthermore, integrating RTS expansions with environmental factors like temperature and precipitation can aid in pinpointing the elements that drive the onset and growth of RTSs (Hayes et al., 2022; Lewkowicz & Way, 2019; Luo et al., 2022; D. Yang, Qiu, et al., 2023). Knowledge about changing boundaries of RTSs is necessary for physical models to accurately estimate and forecast the amounts of carbon, water, and sediment produced during the formation and evolution of RTSs, which is essential for the assessment of subsequent environmental influence (Bernhard et al., 2022; Turner et al., 2021).

### 1.1. Challenges in Generalization of Classic Segmentation Models

Nonetheless, the high cost of visual interpretation and an absence of transferable and automated methods make it challenging to construct multi-annual RTS inventories (e.g., Luo et al., 2022; Xia et al., 2022). Previous studies have shown that deep learning models trained on small-area, single-temporal RTS training data sets exhibit significant performance degradation when the models were directly applied to other areas or time (Huang et al., 2023; Witharana et al., 2022; Xia et al., 2022; Y. Yang, Rogers, et al., 2023).

Such generalization challenge is common when deep learning models are applied to remote sensing images (Ma et al., 2024). Deep learning models typically apply nonlinear transformations to map input data into features that exhibit distinct data distribution patterns (i.e., domains) in high-dimensional space. This enables the models to establish specific criteria to accomplish the tasks. However, when models trained on one data set (the source domain) are directly applied to a new data set (the target domain), discrepancies between the two domains, referred to domain shift, may lead to the failure of the original model's criteria.

The domain shift in remote sensing images has multiple triggers (Ma et al., 2024; Peng et al., 2022; Tuia et al., 2016): (a) Different sensors: variations in sensors may lead to reflectance shift (Wang et al., 2022). For instance, Landsat-8 and Sentinel-2 are often combined to densify the time epochs (e.g., Chen et al., 2021; Pan et al., 2021). However, even for the same area, the observations from the two sensors exhibit significant domain shifts due to differences in central wavelengths, bandwidths, and spatial resolutions (Wu et al., 2022). (b) Different spatial locations: different ground features in various areas correspond to distinct spectral reflectance values. Even for the same feature, variations in its properties across different locations can lead to differences in reflectance. This has become a key bottleneck for large-scale mapping tasks, representative examples include crop type classification (S. Wang, Guan, et al., 2023; Y. Wang, Feng, et al., 2023) and land use mapping (Liu & Li, 2014). (c) Different acquisition time: the ever changing of ground features can also lead to variant reflectance. Crops with seasonal phenological changes (Y. Wang, Feng, et al., 2023) and disasters requiring real-time monitoring (Huang et al., 2021; Islam et al., 2020) are two common targets frequently impacted by multi-temporal domain shifts. (d) Different acquisition conditions: acquisition conditions, including changes in viewing angle, illumination, and weather, can result in variations in reflectance (Matasci et al., 2015).

For multi-temporal RTS mapping over Qinghai-Tibet Plateau, generalization challenges typically arise due to the latter two reasons, that is, differences in acquisition time and conditions. The appearance of RTSs vary considerably over time, as new ones emerge, and existing ones grow larger. The surrounding environment also changes notably throughout seasons, showcasing lush vegetation in summer and snow cover in autumn and winter

(Huang et al., 2021). Additionally, the weather over the Qinghai-Tibet Plateau makes the conditions for acquiring remote sensing images vary greatly, leading to noticeable differences in color tone and texture.

### 1.2. Domain Adaptation for Generalization Improvement

Fortunately, the source domain and target domains still have some commonalities, such as similar mapping targets or spectral bands. Based on these commonalities, Domain Adaptation (DA) techniques can establish connections between the source and target domains, thereby inheriting prior knowledge from the source domain and transferring it to the target domain (Ma et al., 2024; Tuia et al., 2016). The classic DA method, fine-tuning, establishes connections by adapting the models' parameters using a limited number of training samples from the target domain (S. Wang, Guan, et al., 2023). Although proven effective (Zhang et al., 2022; Zhao et al., 2021), fine-tuning largely depends on labels from the target domain.

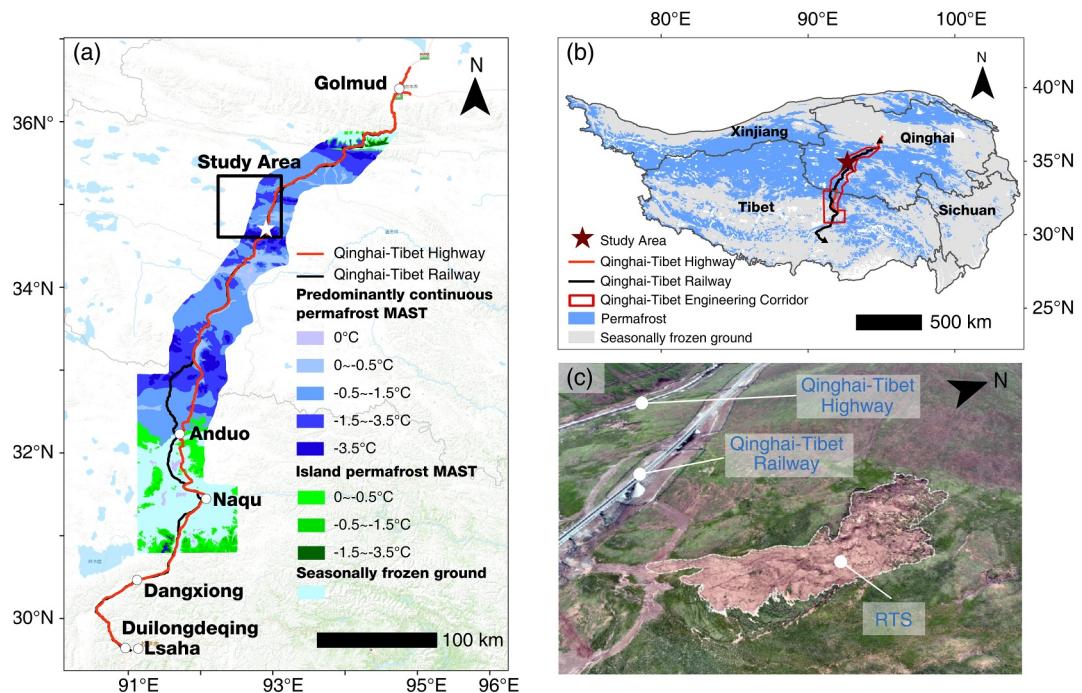
To eliminate the labeling cost, the Unsupervised Domain Adaptation (UDA) has emerged with various methods to directly align the source and target domains (Ma et al., 2024). According to the alignment strategies, the UDA can be categorized into transformation-based methods, discrepancy-based methods, adversarial-based methods, and generative-based methods (Wang & Deng, 2018). The first two methods typically align the source and target domains by either transforming them into a common feature space or minimizing their distribution discrepancy (Othman et al., 2017; Pan et al., 2010). However, both methods involve complex processes for transformation or discrepancy measurements (Chadha & Andreopoulos, 2019). In contrast, adversarial-based and generative-based UDA approaches offer a more straightforward and effective method for domain alignment. Adversarial training, exemplified by Domain Adversarial Neural Networks (DANN), trains a feature extractor to map both source and target domains into domain-invariant features by competing with a domain classifier (Ganin et al., 2016). Generative-based UDA methods, such as GAN, employ generators to transform target domain data into source domain data (Goodfellow et al., 2016). These two UDA methods can effectively improve the generalization, leading to their swift rise in popularity within the field of remote sensing (e.g., Li et al., 2021; Su et al., 2023).

Although DA methods have proven effective in addressing domain shift problems which lead to generalization challenges, they have rarely been systematically applied in multi-temporal RTS mapping, and their potential to improve model generalization remains largely unexplored. Here we aim to systematically apply DA methods to map the dynamically changing margins of RTSs on a multi-annual timescale with the least cost of labels. We started from the DeepLabv3+ model trained solely on single-temporal data, then used the trained model to map the RTSs in other target years and provided quantitative evidence on its limited generalization ability. Next, we proposed three DA methods from different categories to improve the model's temporal generalization ability. In the realm of the supervised DA, we evaluated the efficacy of classic fine-tuning method; In the field of UDA, we were the first to introduce UDA for RTS mapping and developed an Image Style Transfer Domain Adaptation (ISTDA) framework based on CycleGAN. Additionally, we proposed a novel adversarial-based Tversky Adversarial Domain Adaptation (TADA) network, leveraging the DANN framework. We also quantitatively assessed and compared the performance of our DA methods in multi-temporal RTS mapping.

## 2. Study Area and Data Sets

### 2.1. Study Area

Beiluhe region ( $34.65^{\circ}$ – $35.38^{\circ}$  N,  $92.19^{\circ}$ – $93.12^{\circ}$  E) is in the northwestern corner of Qinghai-Tibet Engineering Corridor covering an area of  $6,965 \text{ km}^2$  and has an elevation between 4,487 and 5,552 m (Figures 1a and 1b). With a low annual average temperature of  $-4.8^{\circ}\text{C}$ , Beiluhe region is characterized by extensive ice-rich continuous permafrost. Approximately 70% of the region is underlain by permafrost with volumetric ice contents larger than 30%, and 20% of the region has an ice contents surpassing 50%. The thickness of the permafrost varies between 20 and 80 m, in which the active layer has a thickness of 1.5–2.0 m (Lin et al., 2015; Luo et al., 2015). In the context of global change, the ice-rich permafrost degradation in Beiluhe region has resulted in severe thermokarst disasters including RTSs (Luo et al., 2019; Yu et al., 2016).



**Figure 1.** Beiluhe study area on Qinghai-Tibet Plateau. (a) Qinghai-Tibet Engineering Corridor. The black box shows the geographic location of the study area. The permafrost distribution map is from Tong et al. (2011). (b) Permafrost distribution in Qinghai-Tibet Plateau (Cheng et al., 2011). The red star marks the location of the study area. (c) Image of a Retrogressive Thaw Slump (RTS) near the Qinghai-Tibet Railway and Highway within the study area (Captured on 12 July 2023). White star in panel (a) shows the geographic location of this RTS.

## 2.2. Data Sets

### 2.2.1. PlanetScope Optical Images

The 3-m-resolution PlanetScope CubeSat optical images have four bands (i.e., blue, green, red, and near-infrared), and we applied blue, green, and red bands in our model. These images have been orthorectified and radiometrically corrected, so the pixel values represent the surface reflectance (Planet Team, 2017).

To systematically assess the improvement of temporal generalization ability introduced by different DA methods, we considered images collected in individual years from 2019 to 2022. Due to snowpack and cloud covers, we had to apply multiple image tiles to mosaic an image for each year to fully cover the study area. From July to August, RTSs are relatively active, and stand out in contrast to lush vegetation, thus easier to identify. However, the images acquired from July to August in 2020 are often covered by clouds. Therefore, we used the images mostly from July to August, but opted to October when cloud-free images were unavailable in July and August (Table 1).

### 2.2.2. RTS Ground Truth Boundaries

RTS ground truth boundaries from 2019 to 2022 were delineated referring to the PlanetScope optical images. We verified the boundary polygons with Google Earth Pro historical imagery and 0.5-m-resolution Esri Wayback

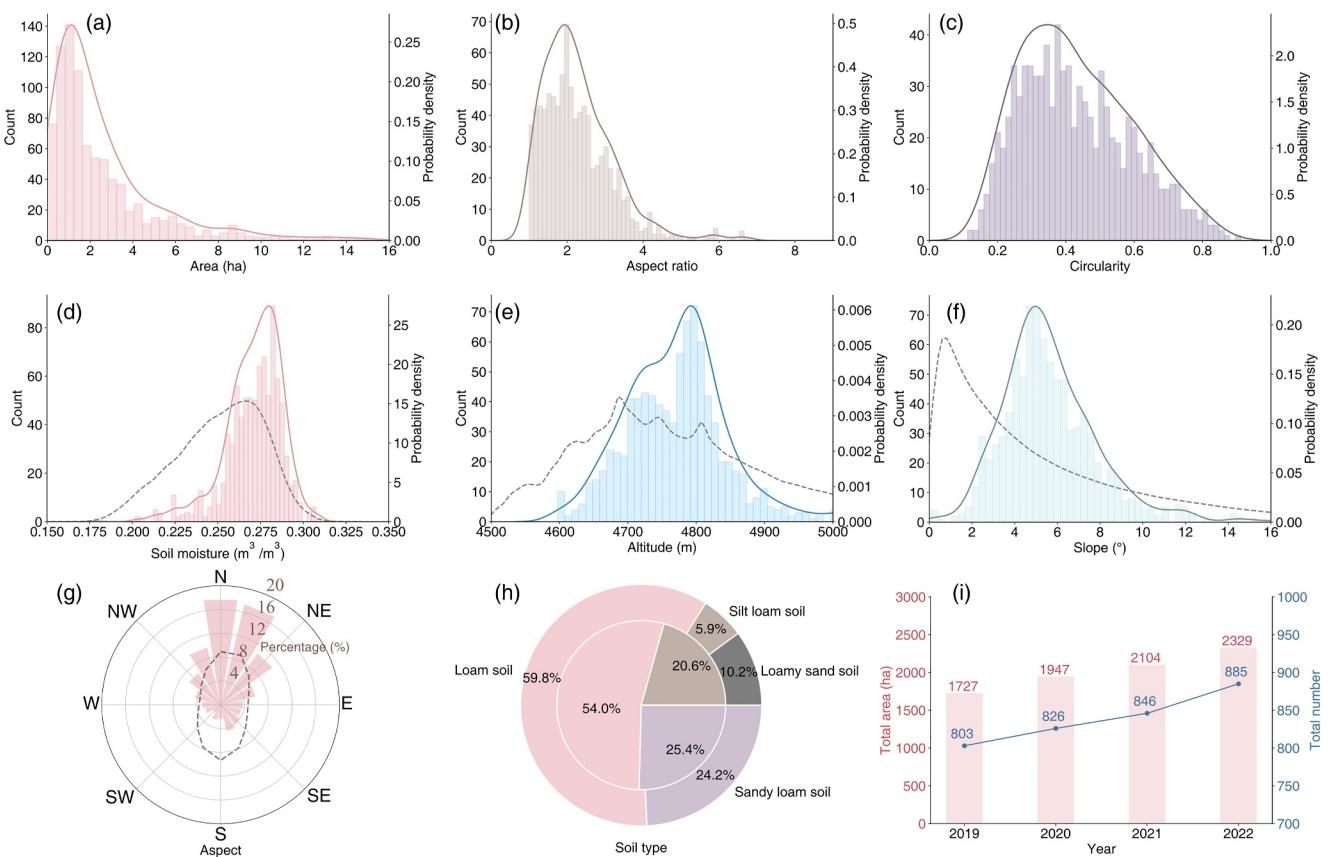
Imagery. We obtained 803, 826, 846, and 885 RTS polygons as the ground truth in 2019, 2020, 2021, and 2022, respectively (Figure 2a). In 2019, ~80% of ground truth was used to generate positive samples in the training data sets for RTS mapping. From 2020 to 2022, the RTS polygons were applied to assess the improvement of our DA methods compared to a direct employment of deep learning model trained in 2019.

The results reveal that RTSs exhibit significant variations in individual areas and an overall elongated shape (Figures 2a–2c). Additionally, the RTSs in the Beiluhe region are highly active, with new slumps emerging and existing ones

**Table 1**

*The Acquisition Dates of PlanetScope Optical Images From 2019 to 2022*

| Year | Acquisition dates     |
|------|-----------------------|
| 2019 | 27 and 30 July        |
| 2020 | 10, 11 and 12 October |
| 2021 | 11 and 12 July        |
| 2022 | 7 July and 1 August   |



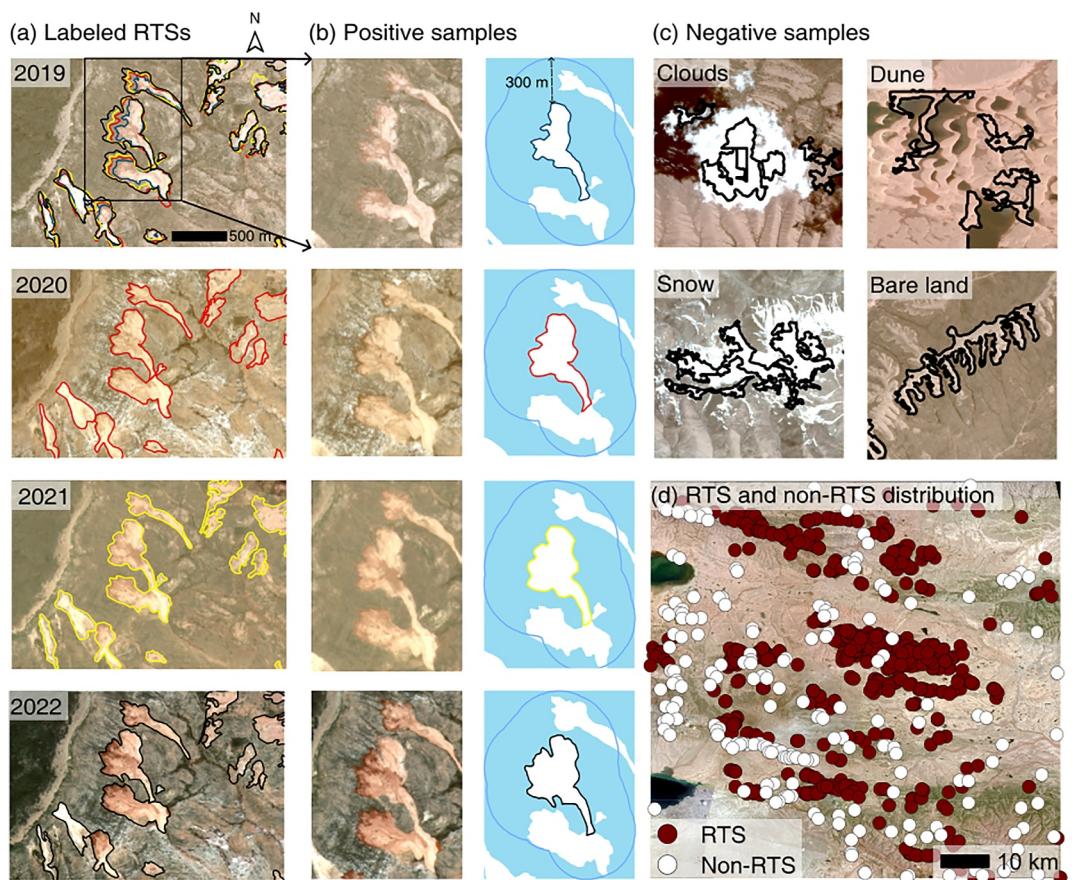
**Figure 2.** Geometric characteristics, influencing factors and dynamic change trends of Retrogressive Thaw Slumps (RTSs) in Beiluhe region. The solid line represents the probability density curve of RTSs and the dashed line represents the probability density curve of the entire study area. In (h), the inner circle represents the soil type proportion of RTSs, while the outer circle represents the soil type proportion of entire study area. The topography data are generated from 30m-resolution Copernicus Global Digital Elevation Model (European Space Agency, 2021). The soil moisture and soil type data are from Liu and Zhang (2022) and Shangguan et al. (2023), respectively.

expanding. The number of RTSs increased from 803 to 885, while the affected area grew from 1,727 ha to 2,329 ha (Figure 2i). Most RTSs in the Beiluhe region are located on north-facing gentle slopes at moderate altitudes, characterized by loam soil with high moisture. At lower altitudes, the underground ice layer is insufficiently thick, while at higher altitudes, the ice layer remains too stable to melt, both obstructing the development of RTSs (Figure 2e). Groundwater trapped by mild slopes can form underground ice and trigger active layer detachment failures by increasing pore pressure. In contrast, steep slopes hinder the convergence of groundwater (Lewkowicz & Harris, 2005; Luo et al., 2019). In addition, moderate slopes provide sufficient gravitational potential for sliding of active layer compared to small slopes (Figure 2f). The active layer on north-facing slopes is thinner due to less solar radiation, rendering it susceptible to be removed and form RTSs (Luo et al., 2022; Xia et al., 2022) (Figure 2g). Regions with higher soil moisture content can form the underground ice more easily (O'Neill & Burn, 2012) (Figure 2d). A higher silt fraction of loam soil implies a higher frost susceptibility, leading to elevated ice levels that are particularly responsive to temperature fluctuations (Gilbert et al., 2016; Lin et al., 2020) (Figure 2h).

### 3. Bottleneck of Classic Segmentation Model

#### 3.1. DeepLabv3+ Semantic Segmentation Model

We used DeepLabv3+ semantic segmentation model (Chen et al., 2018) to construct RTS mapping model trained on single-year data (i.e., 2019 ground truth). In the study, we aim to classify each pixel in PlanetScope images into two classes, that is, RTSs and non-RTSs.



**Figure 3.** The retrogressive thaw slump (RTS) ground truth and generated training data sets. (a) Labeled RTS ground truth from 2019 to 2022, respectively. (b) Positive training samples. (c) Negative training samples. (d) Distribution map of RTS ground truth and selected non-RTS polygons in 2019.

When generating the training data sets, in addition to deriving positive samples from RTS ground truth, we also incorporated non-RTS polygons to obtain negative samples to depress the occurrence of false positives. To identify representative non-RTS polygons that the segmentation model might misclassify as RTSs, we initially ran the DeepLabv3+ model using only positive samples and selected the non-RTS polygons from the predicted false positives. The selected non-RTS polygons primarily include clouds, snow, dunes, and bare land, which exhibit similar morphological or color characteristics to RTSs (Figure 3c). After multiple experiments, we noted that the model performed best when the ratio between RTS ground truth and non-RTS polygons was approximately 4:1. Therefore, in 2019, we selected 200 non-RTS polygons to generate negative samples. To best utilize the surrounding environment information, we considered the minimum bounding box with 300-m buffers to the RTSs as the extent of training images and labels (Figure 3b). To enhance training data diversity, data augmentation methods including blurring, flipping, and rotating to both training and label images were applied (Huang et al., 2020).

During the training stage, we performed fine-tuning on the pre-trained Deeplabv3+ model to inherit prior knowledge and used MobileNetV2 as the backbone architecture due to its high efficiency. During the training, we adopted an initial learning rate of 0.007, a learning rate decay of  $10^{-4}$ , and a batch size of 8 to train the model for 250 epochs.

### 3.2. Evaluation Metrics

In classic semantic segmentation tasks, model performance is typically evaluated using pixel-based metrics. However, these metrics often prioritize larger targets. This means that even if models miss smaller targets, pixel-based metrics can still appear high, since small targets account for only a minor portion of the overall pixel count.

**Table 2**

*Quantitative Performance of DeepLabv3+ Trained on the Source Domain (2019) in Both the Source and Target Domains (2020–2022)*

| IoU threshold metrics (%) | 0      |           |       | 0.4    |           |       | 0.8    |           |       |
|---------------------------|--------|-----------|-------|--------|-----------|-------|--------|-----------|-------|
|                           | Recall | Precision | F1    | Recall | Precision | F1    | Recall | Precision | F1    |
| 2019                      | 87.67  | 80.46     | 83.91 | 84.31  | 77.37     | 80.69 | 61.15  | 56.11     | 58.52 |
| 2020                      | 62.71  | 34.19     | 44.25 | 48.31  | 26.34     | 34.09 | 16.34  | 8.91      | 11.53 |
| 2021                      | 54.61  | 47.34     | 50.71 | 43.26  | 37.50     | 40.18 | 11.23  | 9.73      | 10.43 |
| 2022                      | 66.89  | 37.54     | 48.09 | 53.11  | 29.80     | 38.18 | 13.67  | 7.67      | 9.83  |

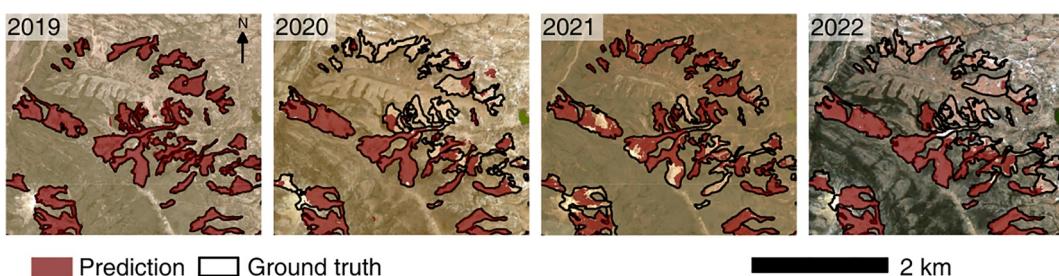
Given the wide variation in RTS areas (Figure 2a), pixel-based metrics may fail to accurately assess the model's performance in mapping small RTSs, which account for a significant proportion in most RTSs.

To overcome this limitation, we employed a polygon-based F1 score, commonly used in previous RTS mapping studies (e.g., Huang et al., 2020; Rodenhizer et al., 2024; Y. Yang, Rogers, et al., 2023). Polygon-based metrics treat each RTS as an individual unit, calculating the Intersection over Union (IoU) between each RTS and its corresponding mapped polygon. An IoU greater than 0 indicates partial mapping of the RTSs, while an IoU of 1 signifies perfect alignment with the ground truth boundary. A mapped polygon is classified as a true positive if its IoU exceeds a specified threshold (e.g., 0.5); otherwise, it is labeled as a false positive. Therefore, we can derive polygon-based recall, precision, and F1 score metrics, which assess the mapping efficiency of multi-scale RTSs equally. For simplification, we used F1 to represent polygon-based F1 score hereafter.

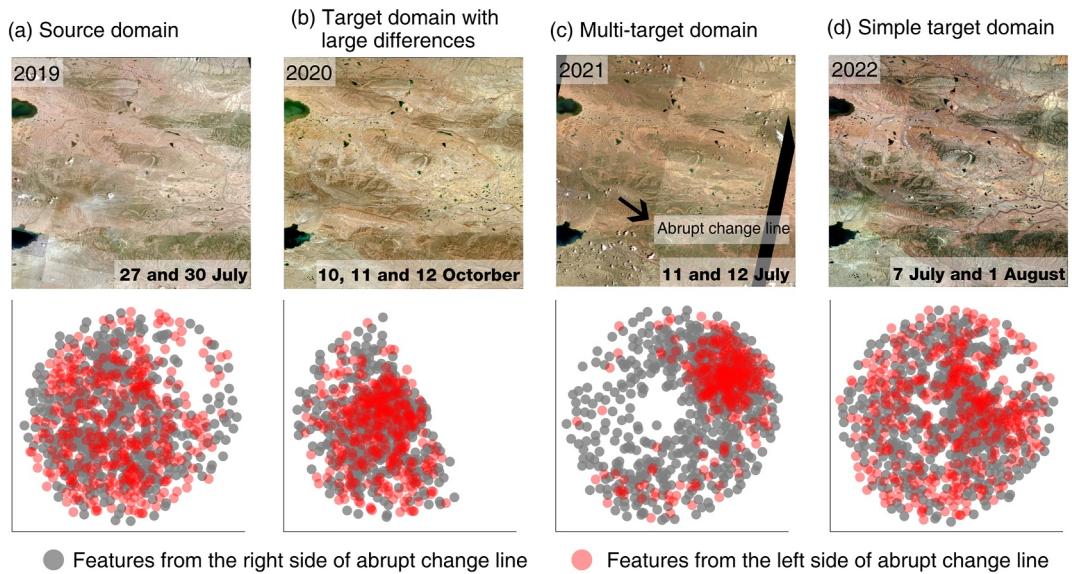
Additionally, the IoU threshold can be adjusted to meet specific mapping objectives. Setting lower IoU thresholds can capture a more comprehensive RTS inventory with fewer RTSs missed. Although mapped polygons with lower IoU scores may not align perfectly with the ground truth, they still offer valuable information about RTS locations and reduce the likelihood of overlooking RTSs. These locations can help recover RTS boundaries in subsequent work. In contrast, high IoU threshold emphasize the completeness of individual RTS. While this approach may increase the risk of missing some RTSs, it ensures greater reliability in the mapped results. For our evaluation, we selected three IoU thresholds: 0, 0.4, and 0.8. The threshold of 0 evaluates the model's ability to capture RTSs' locations, while 0.8 focuses on the completeness of individual RTS mappings. The 0.4 threshold balances both aspects.

### 3.3. Model's Generalization Issue

To evaluate the classic model's temporal generalization ability, we applied the trained model to multi-temporal images collected from 2019 to 2022. On the 2019 PlanetScope images, the DeepLabv3+ model accurately delineated RTS boundaries compared with the ground truth from visual interpretation. The F1 scores were 83.91%, 80.69%, and 58.52% at IoU thresholds of 0, 0.4, and 0.8, respectively, with high precision and recall. However, when the model trained on the 2019 data set was directly applied to the 2020, 2021, and 2022 images, its performance declined significantly (Table 2). The predicted polygons did not align well with the ground truth, with an appreciable number of missing RTSs and false positives exist (Figure 4). In 2020, 2021, and 2022, at an



**Figure 4.** Visualized performance of the classic model trained on the source domain (2019) in both the source and target domains (2020–2022).



**Figure 5.** The mosaiced images from 2019 to 2022 and quantitative visualization of domain shifts between source and target domains. The acquisition dates are annotated at the bottom-left corner of the images for each year. (a) Source domain (2019). (b) Target domain with distinct differences from the source domain (2020). (c) Multi-target domain (2021). (d) Simple target domain (2022).

IoU threshold of 0, the number of false positives reached 997, 514, and 985, respectively, accounting for 65.81%, 52.66%, and 62.46% of the predicted results.

### 3.4. Complex Target Domain

This section explains the dramatic performance degradation through a detailed analysis to the images and a quantitative visualization of the domain shift between the source year (2019) and the target years (2020–2022). We used t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of the mapped features from high-dimensional space to two dimensions, allowing for the visualization of data distribution patterns in the source and target domains. The results show clear differences between the source and target domains (Figure 5). The common causes of the domain shifts are dynamic changes of RTSs and image acquisition condition variation between years. For each year, new RTSs emerge and existing RTSs keep expanding upward rapidly (Figures 2i and 3a). And the images in 2020–2022 exhibit variations in texture and color tone compared to those in 2019 due to different image acquisition conditions.

Additionally, several unique factors give rise to three distinct target domain scenarios: (a) Simple target domain: the RTSs in 2022 share similar lush vegetation surroundings with 2019 due to the close acquisition seasons. The mosaiced images are also of high quality, with a high level of consistency and uniformity. As a result, the data distribution pattern of 2022 closely resembles that of 2019, forming a simple target domain (Figure 5d). (b) Target domain with large differences from the source domain: While the images of other years were collected in summer, the mosaiced image in 2020 was collected in October. The RTSs were covered by snow and the surrounding vegetation was sparse in October, contrasting sharply with the lush vegetation in the summer, resulting in a large difference in data distributions between 2019 and 2020 (Figure 5b). (c) Multi-target domain: Aside from the year-to-year acquisition differences, the mosaiced images from 2021 exhibit noticeable inconsistency, with a distinct color tone differences due to varying acquisition conditions between adjacent scenes. We divided the image into two sides based on the abrupt change line, and processed both sides separately using t-SNE. The results reveal each side exhibit distinct data distributions, forming a multi-target domain (Figure 5c). To rule out the possibility that the differences between the two sides were caused by other factors, we also performed t-SNE analysis on the features from both sides of the remaining years (i.e., 2019, 2020 and 2022). Our results demonstrate that the features of other years did not show significant differences between these two sides.

## 4. Domain Adaptation Methods

We incorporated three representative domain adaptation approaches from different categories to address the challenges of generalization in multi-temporal RTS mapping tasks.

### 4.1. Fine-Tuning Based Domain Adaptive Segmentation

Relying on the trained model in the source domain, we supplemented a small amount of training data from target domain to adapt the model parameters. Convolution neural networks (CNNs) typically consist of multiple layers. The lower layers are designed to identify low-level features like edges, corners, and textures, while the higher layers are built upon low-level features and capture more abstract features of a certain pattern. We froze the parameters in the lower layers and retrained the parameters in higher layers using the training data from the target domain. Consequently, the fine-tuned model can fully utilize the prior knowledge obtained from source domain and incorporate new knowledge specified in the target domain, leading to a better transfer performance.

In this study, we randomly selected 50, 100, and 150 RTS ground truth samples from the target domains (2020–2022) to generate positive training data. Before fine-tuning, the model was run with only positive samples to obtain the corresponding negative samples. Following the previously established optimal ratio of 4:1 between ground truth and negative polygons, we incorporated 12, 24, and 36 negative polygons into the 50, 100, and 150 ground truth samples, respectively, to create the final training data sets. Using these data sets, we fine-tuned the model and evaluated its temporal generalization ability.

### 4.2. Image Style Transfer Domain Adaptation Framework (ISTDA)

Instead of fine-tuning the model with labels from the target domain, CycleGAN directly transforms target domain images into new images that follow the source domain's data distribution, without incurring any label costs. The transformation is accomplished through an adversarial training process between a generator and a discriminator. The generator aims to produce fake images to fool the discriminator, while the discriminator aims to distinguish the real images from the synthesized ones from the generator. During the process of training, the generator and discriminator engage in a battle until the discriminator can no longer distinguish between the real and the generated images (Goodfellow et al., 2016; Zhu et al., 2017).

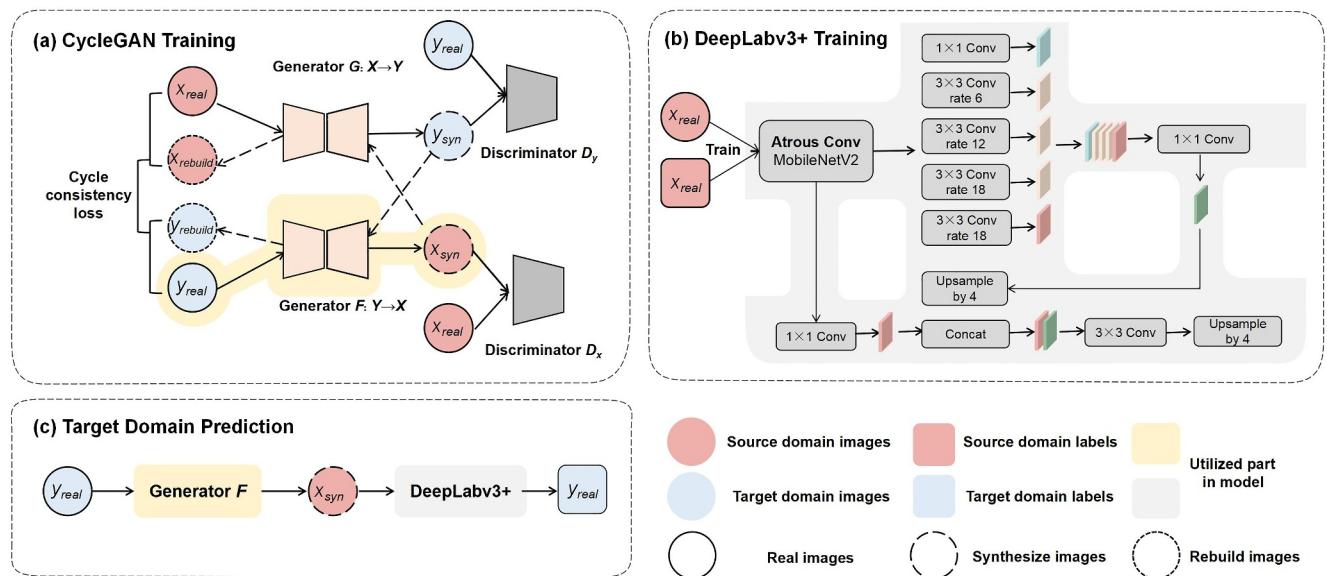
As a step further, we proposed ISTDA based on CycleGAN. We applied a 1-km buffer to 2019 RTS boundaries and cropped the optical images from both source domain (2019) and target domains (2020–2022) into sub-images referring to the bounding box of the buffer zone. We further divided the sub-images into patches in size of 300 by 300 pixels and maintained an overlap of 90 pixels between adjacent patches.

In the CycleGAN model, we adopted a learning rate of  $10^{-4}$ , an identity loss of 1, a batch size of 8, and trained the model for 400 epochs. The identity loss helps to maintain the consistency between input images and the synthesized ones. The trained generator enables us to transform the visual characteristics of target-domain (2020–2022) images into that of the source-domain (2019) images; nonetheless, the RTS boundaries are unchanged. The synthesized images were finally ingested into DeepLabv3+ model trained on 2019 data sets to predict the RTS boundaries in other years (Figure 6).

### 4.3. Tversky Adversarial Domain Adaptation (TADA)

While CycleGAN addresses domain shift by transforming the target domain to match the source domain's distribution at the image level, DANN tackles this challenge by mapping features from both domains into the same data distribution in a high-dimensional space (Ganin et al., 2016). The DANN framework avoids labeling cost through a structure that consists of a feature extractor, a domain classifier, and a label predictor. During training, the feature extractor competes with the domain classifier, striving to learn domain-invariant features while the classifier attempts to distinguish between the source and target domains. A Gradient Reversal Layer (GRL) assists the feature extractor by reversing the gradients of the domain classifier. Finally, the label predictor uses the extracted features to generate prediction results.

Based on the DANN framework, we proposed a novel network named TADA. The proposed TADA network takes a part (before the concatenation layer of two feature layers) of the pre-trained DeepLabv3+ model as the feature extractor. This part contains an encoder with Atrous Spatial Pyramid Pooling and a decoder, allowing for



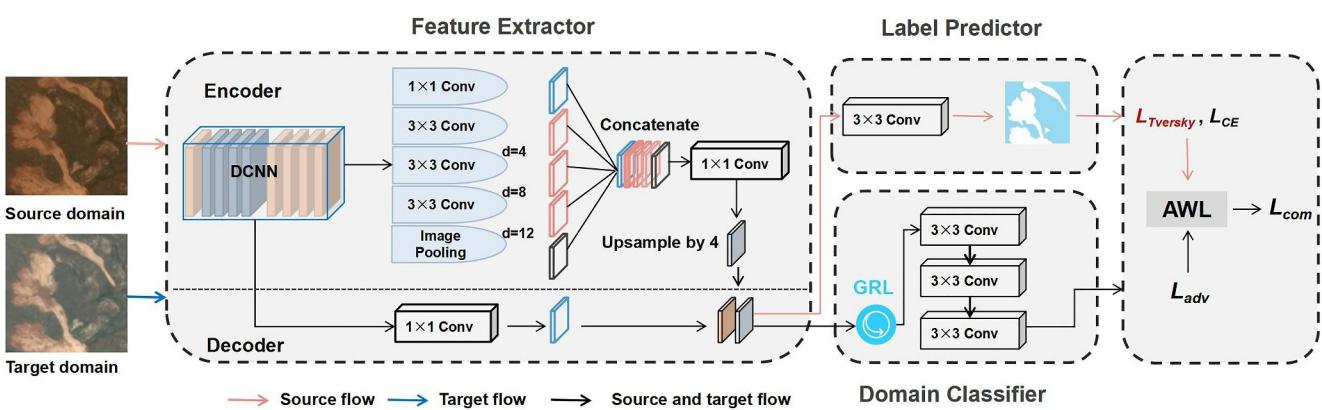
**Figure 6.** The workflow of Image StyleTransfer Domain Adaptation.

the extraction of multi-scale and multi-level information. The remaining section of the Deeplabv3+ model was used as the label predictor. Simultaneously, a domain classifier was connected to the feature extractor via GRL. This domain classifier consists of three  $3 \times 3$  convolutional layers, separating features from the source and target domains apart (Figure 7).

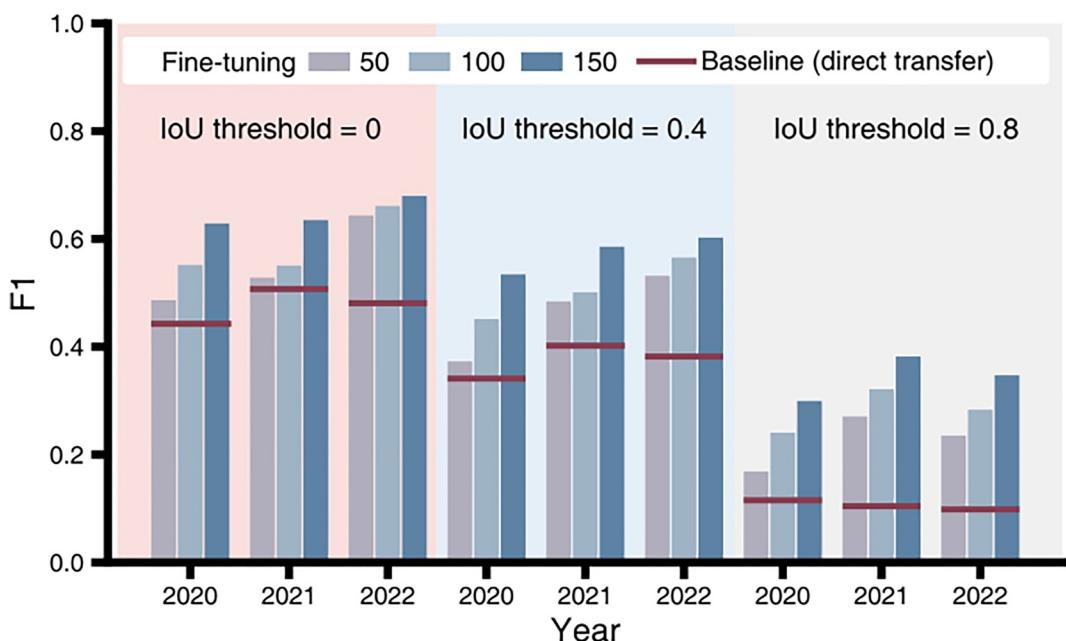
Beyond simply adopting the DANN framework, we further integrated Tversky loss into the label predictor's loss function to mitigate excessive false positives when transferring. In our binary classification task, the Tversky loss is given by

$$\left\{ \begin{array}{l} L_{Tversky}(y, \hat{y}) = \frac{\sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N (y_i \hat{y}_i + \alpha(1 - y_i) \hat{y}_i + \beta y_i (1 - \hat{y}_i))} \\ \alpha + \beta = 1 \end{array} \right. \quad (1)$$

where  $y$  represents ground truth labels,  $\hat{y}$  represents the predicted labels, and  $\alpha$  and  $\beta$  are two hyperparameters that modulate the penalties for false positives and false negatives. When  $\alpha > \beta$ , the loss function will suppress the false



**Figure 7.** The architecture of the Tversky Adversarial Domain Adaptation.



**Figure 8.** F1 increment based on fine-tuning based domain adaptive segmentation. Red lines represent F1 results when the trained model from the source domain (2019) is directly applied to the target domain (2020–2022).

positives. After multiple tuning processes, we determined the hyperparameters ( $\alpha = 0.7$ ;  $\beta = 0.3$ ) to achieve optimal performance.

Thereafter, the total loss function comprises three components: the CE loss and the newly incorporated Tversky loss from the label predictor, as well as the adversarial loss from the domain classifier (Figure 7). It becomes a multi-objective optimization task to solve for the weights among these three components. In this multi-objective optimization neural network, the total loss function can be defined as:

$$L_{\text{comb}}(x, y_T, y'_T; \omega_T) = \sum_{\tau \in T} L_\tau(x, y_\tau, y'_\tau; \omega_\tau) \cdot c_\tau \quad (2)$$

where  $\omega_T$  represents trainable parameters in the model unrelated to the loss,  $x$  represents the input images,  $y_T$  represents the ground truth labels,  $y'_T$  represents the model prediction labels, and  $c_\tau$  represents the weights of several loss functions. Instead of manually adjusting these weights, we introduced the Automatic adjustment of Weights for multiple Loss functions (AWL) (Liebel & Körner, 2018). This method considers the weights as learnable parameters to improve the training efficiency and model performance.

## 5. Results

### 5.1. Fine-Tuning Based Domain Adaptive Segmentation

Fine-tuning based domain adaptive segmentation can effectively enhance the temporal generalization ability (Figure 8). Fine-tuning the model with different amounts of training data yields a significant increase in F1 score from 3.19% to 22.08% at an IoU threshold of 0.4. As the amount of training data from the target domain increases, the F1 improvement becomes more pronounced. This is because more training data samples provide more information from the target domain. More importantly, fine-tuning based domain adaptive segmentation can alleviate the performance degradation in complex target domain scenarios involving target domain with large differences from source domain (2020) and multi-target domain (2021).

However, fine-tuning based domain adaptive segmentation struggles to suppress false positives effectively, achieving a low precision of only 39.32%–59.93% at an IoU threshold of 0 (Table 3). This issue is especially severe in the multi-target domain scenario (2021). The challenge arises from the multi-domain scenario with two

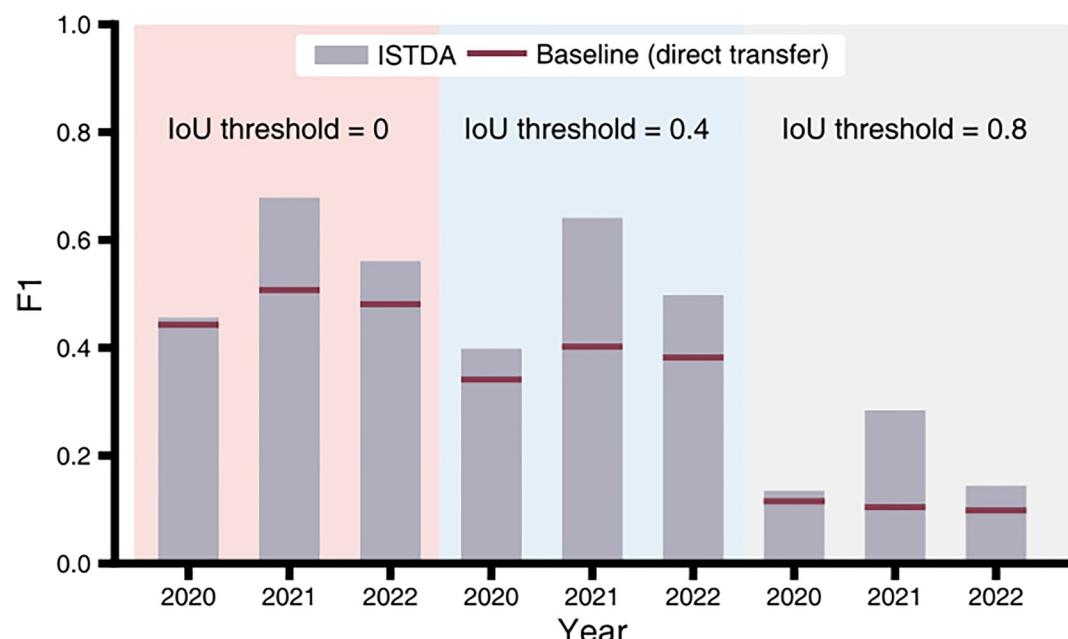
**Table 3***Quantitative Performance of Fine-Tuning Based Domain Adaptive Segmentation*

| Number of RTSs for fine-tuning | IoU threshold<br>Metrics (%) | 0      |           |       | 0.4    |           |       | 0.8    |           |       |
|--------------------------------|------------------------------|--------|-----------|-------|--------|-----------|-------|--------|-----------|-------|
|                                |                              | Recall | Precision | F1    | Recall | Precision | F1    | Recall | Precision | F1    |
| 50                             | 2020                         | 50.24  | 47.16     | 48.65 | 38.50  | 36.14     | 37.28 | 17.43  | 16.36     | 16.88 |
|                                | 2021                         | 80.50  | 39.32     | 52.83 | 73.76  | 36.03     | 48.41 | 41.25  | 20.15     | 27.08 |
|                                | 2022                         | 71.41  | 58.57     | 64.36 | 58.98  | 48.38     | 53.16 | 26.10  | 21.41     | 23.52 |
| 100                            | 2020                         | 66.59  | 47.09     | 55.17 | 54.48  | 38.53     | 45.14 | 29.06  | 20.55     | 24.07 |
|                                | 2021                         | 82.86  | 41.24     | 55.07 | 73.76  | 36.03     | 48.41 | 41.25  | 20.15     | 27.08 |
|                                | 2022                         | 77.29  | 57.77     | 66.12 | 66.10  | 49.41     | 56.55 | 33.11  | 24.75     | 28.32 |
| 150                            | 2020                         | 66.10  | 59.93     | 62.87 | 56.17  | 50.93     | 53.43 | 31.48  | 28.54     | 29.94 |
|                                | 2021                         | 82.15  | 51.75     | 63.50 | 75.77  | 47.73     | 58.57 | 49.41  | 31.12     | 38.19 |
|                                | 2022                         | 78.76  | 59.78     | 67.97 | 69.83  | 53.00     | 60.26 | 40.23  | 30.53     | 34.71 |

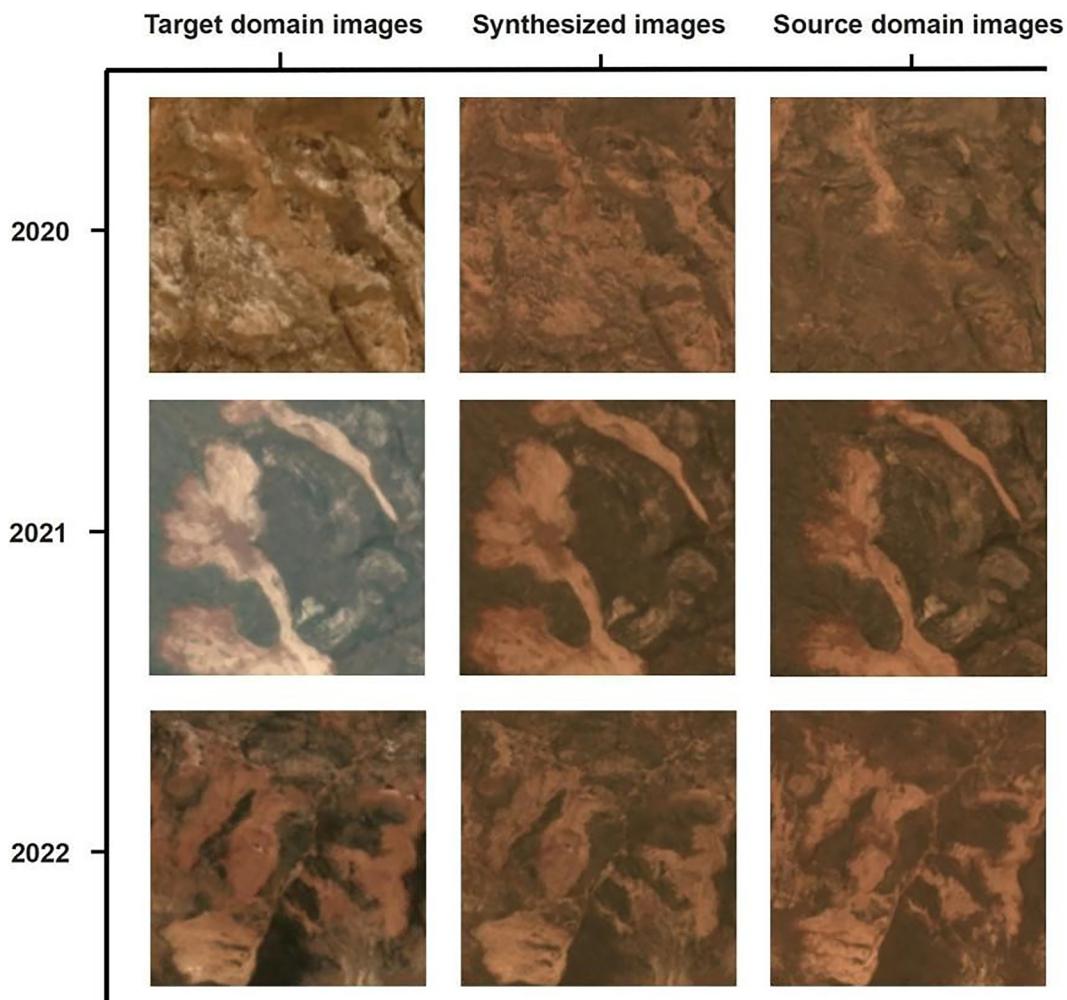
distinct target domains, leading to more complex and varied RTS-like features. The limited number of negative samples are insufficient to cover these features, thereby increasing false positives.

## 5.2. Image Style Transfer Domain Adaptation Framework

When the differences between the source and target domains are small, such as in the case of 2021 and 2022, ISTDA can effectively enhance the temporal generalization ability (Figure 9). In comparison between the 2 years, ISTDA performed better in 2021 than in 2022 with F1 improved by 17.14%, 23.91%, and 17.96% when IoU thresholds of 0, 0.4, and 0.8 are applied, respectively. This is because CycleGAN can remarkably transfer the style of 2021 images to that of source images collected in 2019 while maintaining RTS boundaries (Figure 10). On the other hand, the improvement in temporal generalization ability in 2022 is less pronounced with F1 improved by 7.97%, 11.59%, and 4.57% when applying IoU thresholds of 0, 0.4, and 0.8, respectively (Figure 9). This is because 2022 images have already shared similar style to those collected in 2019 (Figure 10), and thus the potential for further improvement in 2022 scenario through ISTDA is limited.



**Figure 9.** F1 increment based on Image StyleTransfer Domain Adaptation. Red lines represent F1 results when directly applying the trained model from the source domain (2019) to the target domain (2020–2022).

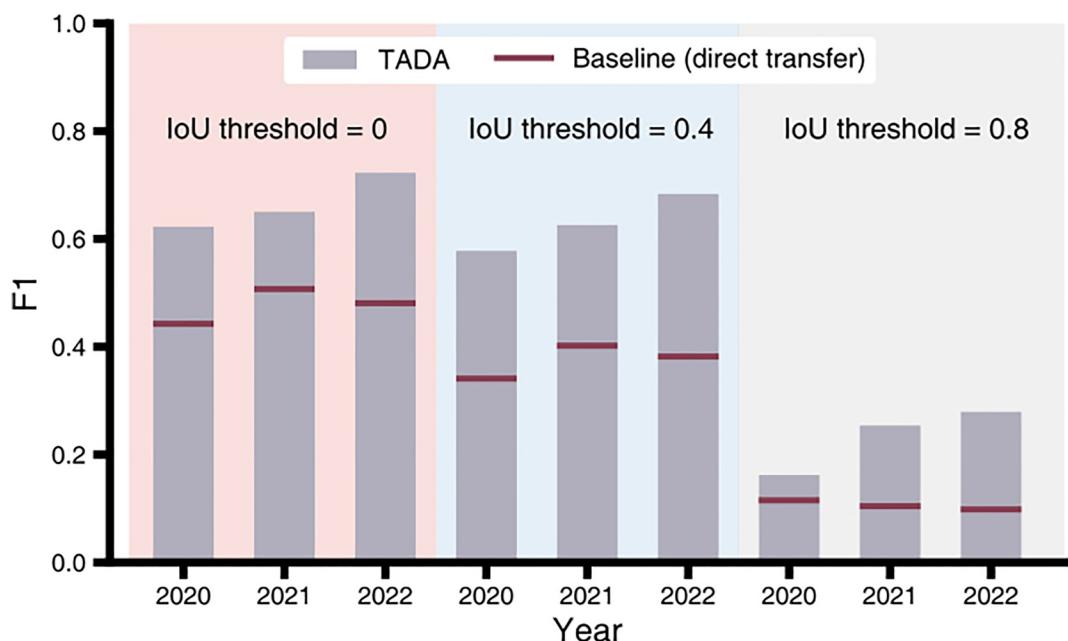


**Figure 10.** PlanetScope images from the target domain (2020–2022), the synthesized images generated by CycleGAN, and the images from the source domain (2019).

When differences exist between the source and target domains are relatively large (e.g., in 2020), ISTDA may not substantially enhance the temporal generalization ability of the RTS mapping model, resulting in a high rate of false positives (Figure 9). In 2020, ISTDA's precision remains low at only 35.71%, 31.19%, and 10.56% for IoU thresholds of 0, 0.4, and 0.8, respectively (Table 4). This limitation arises because the target domain, differing greatly from the source domain, includes new RTS-like features such as snow cover which are absent in 2019. Figure 10 illustrates misleading synthesized images generated by CycleGAN, where snow-covered areas are transformed to closely resemble the color and texture of RTSs, causing the model to misclassify RTSs in these synthesized images.

**Table 4**  
*Quantitative Performance of Image StyleTransfer Domain Adaptation*

| IoU threshold metrics (%) | 0      |           |       | 0.4    |           |       | 0.8    |           |       |
|---------------------------|--------|-----------|-------|--------|-----------|-------|--------|-----------|-------|
|                           | Recall | Precision | F1    | Recall | Precision | F1    | Recall | Precision | F1    |
| 2020                      | 63.08  | 35.71     | 45.60 | 55.08  | 31.19     | 39.82 | 18.64  | 10.56     | 13.48 |
| 2021                      | 74.59  | 62.23     | 67.85 | 70.45  | 58.78     | 64.09 | 31.21  | 26.04     | 28.39 |
| 2022                      | 69.49  | 46.98     | 56.06 | 61.69  | 41.71     | 49.77 | 17.85  | 12.07     | 14.40 |



**Figure 11.** F1 increment based on Tversky Adversarial Domain Adaptation. Red lines represent F1 results when directly applying the trained model from the source domain (2019) to the target domain (2020–2022).

### 5.3. Tversky Adversarial Domain Adaptation

Our experiments show that TADA effectively enhances the temporal generalization ability in complex target domain scenarios. Compared with the direct transfer, the F1 of the TADA increases by 14.32%–24.17% when the IoU threshold is 0. F1 has been improved by 22.40%–30.18% when the IoU threshold is 0.4, and by 4.67%–18.10% when the IoU threshold is 0.8 (Figure 11). TADA demonstrates exceptional recall rates, ranging from 77.24% to 83.62% at an IoU threshold of 0. However, its precision remains relatively low compared to recall rates, between 52.17% and 63.63% at an IoU threshold of 0, leading to a higher number of false positives in prediction results (Table 5).

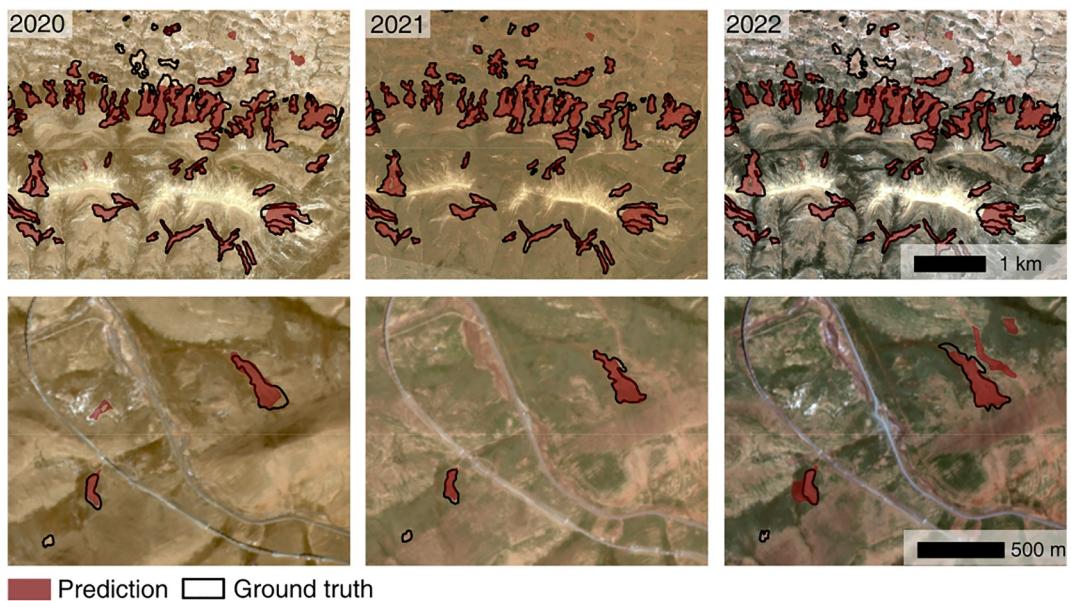
## 6. Discussion

### 6.1. The Necessity of Applying Domain Adaptation

The incorporation of domain adaptation methods in multi-temporal RTS mapping is necessary. Although it may be feasible to manually interpret RTSs or manually verify the prediction results of classic deep learning models in small areas over a few years, these approaches are impractical for larger-extent or long-term applications. For instance, the Qinghai-Tibet Plateau and the Arctic are two regions many RTS have developed, encompassing areas of 2.5 million km<sup>2</sup> and 8 million km<sup>2</sup>, respectively. The rapid initiation and expansion of RTSs in these areas have persisted for decades (Bernhard et al., 2022; Luo et al., 2019). Manual interpretation can cost considerable labor and time. Collecting sufficient RTS ground truth to effectively train the models is also almost unlikely. Besides, directly transferring deep learning models trained on limited data often results in numerous RTSs being

**Table 5**  
*Quantitative Performance of Tversky Adversarial Domain Adaptation*

| IoU threshold | 0           |        |           | 0.4   |        |           | 0.8   |        |           |
|---------------|-------------|--------|-----------|-------|--------|-----------|-------|--------|-----------|
|               | Metrics (%) | Recall | Precision | F1    | Recall | Precision | F1    | Recall | Precision |
| 2020          | 77.24       | 52.17  | 62.27     | 71.67 | 48.41  | 57.78     | 20.10 | 13.57  | 16.20     |
| 2021          | 81.44       | 54.12  | 65.03     | 78.37 | 52.08  | 62.58     | 31.80 | 21.13  | 25.39     |
| 2022          | 83.62       | 63.63  | 72.27     | 79.10 | 60.19  | 68.36     | 32.32 | 24.59  | 27.93     |



**Figure 12.** Prediction results of Tversky Adversarial Domain Adaptation in a Retrogressive Thaw Slump cluster and in a localized area near the Qinghai-Tibet Railway and Highway from 2020 to 2022.

missed and many false positives in prediction results. In our study area, when the IoU threshold is 0.4, only about 40%–50% of RTSs can be correctly mapped, while approximately 70% of the predictions are false positives. On a larger scale, Huang et al. (2023) employed another classic deep learning model, YOLOv4, to identify RTSs in the Arctic. The model trained with 563 RTS ground truth ultimately generated 524,240 predictions—over 200 times the size of the final inventory, which included 2,494 RTSs. Alternatively, domain adaptation methods can significantly enhance the performance of classic models, reduce reliance on RTS ground truth during training, and the cost of post-processing manual inspection for compiling final inventories.

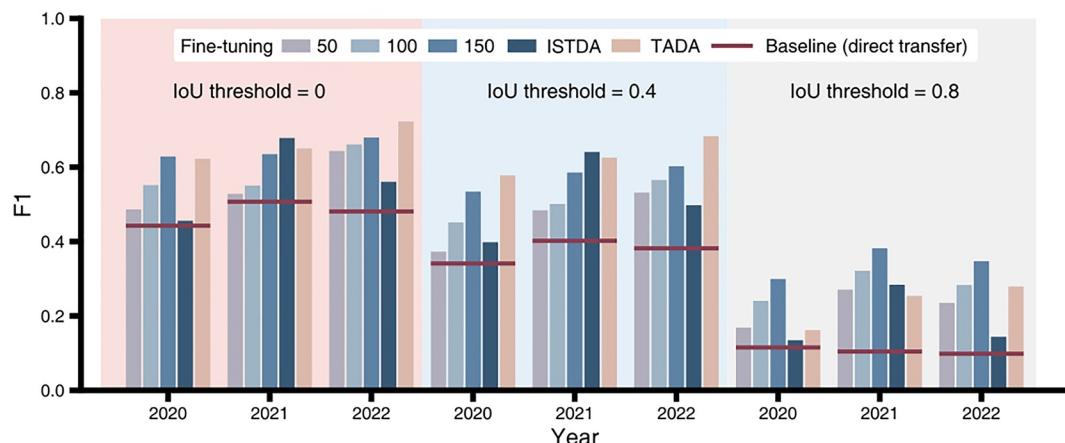
## 6.2. The Applicability of Prediction Results

To assess the applicability of prediction results from the three domain adaptation methods in engineering safety and scientific investigations, we visualized the mapped RTSs near the Qinghai-Tibet Railway and Highway, as well as within an RTS cluster. The visualization results demonstrated the effectiveness of the predictions in qualitatively depicting the distribution, boundaries, and dynamic changes of RTSs. In the RTS cluster, most RTSs were mapped with relatively high alignment to the ground truth; however, some RTSs were still missed, and false positives exist. For the engineering project, two relatively large RTSs were identified, with the dynamic trends of the larger one clearly reflected. For the smaller RTSs, the mapped boundaries overlap well with the ground truth in 2020 and 2021, yet evident errors exist for 2022 (Figure 12).

Although the prediction results successfully depict the general outlines of RTS boundaries and their dynamic changes, they still exhibit a relatively high number of false positives, boundary inaccuracies, and missing RTSs (see Section 6.5). Therefore, the prediction results are insufficient for some quantitative analyses of the retreat rates, released mass, and subsequent environmental impacts. Improving the algorithm can help alleviate this issue. However, no automatic method based on deep learning can achieve perfect alignment between ground truth and prediction results. Therefore, the prediction results can be further refined through manual verification and post-processing methods to achieve these quantitative applications (e.g., Huang et al., 2021; Xia et al., 2022).

## 6.3. Comparison Between Different Domain Adaptation Methods

Here we evaluated and compared three methods in terms of implementation cost and performance to inform model selection for different demands. We compared the F1 results of three domain adaptation methods to assess the temporal generalization ability (Figure 13). At low to moderate IoU thresholds (i.e., 0 and 0.4), two unsupervised methods (TADA and ISTDA) outperform the classic supervised fine-tuning based domain adaptive



**Figure 13.** F1 increment based on three domain adaption methods. Red lines represent F1 results when directly applying the trained model from the source domain (2019) to the target domain (2020–2022).

segmentation. Between the two unsupervised methods, TADA performs better than ISTDA when the target domain is not multi-domain (i.e., 2020 and 2022). Compared to the direct application, the F1 improvements range from 14.32% to 30.18% at IoU thresholds of 0 and 0.4. When the target domain is multi-domain (i.e., 2021), ISTDA performs best, with F1 improvements of 17.14% and 23.91% at IoU thresholds of 0 and 0.4, respectively. In contrast, in the condition of high IoU thresholds, the fine-tuning based domain adaptive segmentation performs the best against ISTDA and TADA, achieving an F1 improvement ranging from 5.35% to 27.76% compared to direct application. When considering label costs, 50 to 150 of RTS boundaries in the target domain is required in fine-tuning. This process is labor intensive and time consuming. However, ISTDA and TADA do not require additional labels from the target domain.

As the results indicate, unsupervised methods have lower completeness in some of the mapped RTSs, but fewer RTSs are entirely missed. In contrast, fine-tuning achieves higher completeness in mapped RTSs, but also result in more entirely missed RTSs. Their contrasting characteristics make them suitable for different tasks. Unsupervised methods can effectively support infrastructure protection. For railways and highways, which are vulnerable to RTSs, these methods enable high-frequency monitoring, reducing the risk of infrastructure damage caused by missed RTSs. Supervised methods may be more suitable for sophisticated tasks requiring high completeness of individual RTS, such as measuring retreat rates and released mass quality.

When constructing comprehensive RTS inventories, supervised methods require ground truth collection for each transfer, whereas unsupervised methods are free from label costs. Besides, the prediction results of both supervised and unsupervised methods do not yield comprehensive RTS inventories due to missed RTSs and false positives. Therefore, further verification to filter out false positives and supplement the missed RTSs is necessary. TADA's recall rate, reaching approximately 80% at an IoU threshold of 0, surpasses that of fine-tuning based methods in most cases. While some mapped polygons may not fully align with the corresponding ground truth, they still offer valuable locational information for most RTSs, thereby reducing the effort needed to identify missed RTSs during verification. Additionally, the precision of fine-tuning based methods remains lower than that of TADA at an IoU threshold of 0, leading to a greater effort required to filter out false positives compared to TADA. Although TADA's completeness for recalled RTSs is lower than that of the fine-tuning based method, recovering boundaries based on known locations is significantly easier than searching for missed RTSs across a large area. Overall, TADA demands less labor and time than fine-tuning based methods during model training and post-verification, making it more suitable for large-scale and long-term multi-temporal RTS inventorying.

#### 6.4. Contribution of This Work

Our work systematically evaluates the performance of domain adaptation methods from different categories in mapping multi-temporal RTSs. In addition to the traditional supervised fine-tuning approach, we proposed two novel methods based on adversarial and generative UDA, respectively. To the best of our knowledge, this is the

first application of UDA methods for mapping multi-temporal RTSs, as well as the first systematic evaluation to domain adaptation techniques from various categories for multi-temporal RTS mapping. The proposed methods are proven effective in improving the performance in multi-temporal RTSs mapping, envisioning the potential of domain adaptation methods in monitoring large-scale and multi-temporal thermokarst and other geomorphological features.

In addition to the technical contributions, our work targets an important yet rarely studied field: the automatic construction of a multi-temporal RTS inventory. Compared to the commonly used manual interpretation, the proposed automatic methods can generate long-term and continuous multi-temporal RTS inventories rapidly. As a step forward from the widely studied single-temporal RTS inventories, the multi-annual RTS inventories support the engineering safety by identifying and monitoring those developed close to the highways and railways. Moreover, it can improve our understanding about RTS initiation and development, and further facilitate the estimation of their environmental impacts.

### 6.5. Limitation and Future Work

The main limitation of this study lies in the considerable number of false positives across all three domain adaptation methods. Among them, TADA demonstrates relatively fewer false positives and maintains stable performance across different target domain scenarios. In contrast, fine-tuning based domain adaptive segmentation and ISTDA struggle in certain scenarios: the former underperforms in multi-target domains, and the latter is less effective when the target domain differs greatly from the source domain.

The considerable number of false positives increases the verification workload when compiling a comprehensive RTS inventory and can be attributed to the extensive, highly dynamic, and variable background environments. Although we included surrounding environments of RTSs in positive samples and RTS-like features in negative samples when training the three domain adaptation methods, these samples cannot encompass all types of RTS-like features due to the significant imbalance between RTSs and the background. Complex target domains may introduce new RTS-like features due to seasonal changes (e.g., snow cover), diverse ground surfaces, and varying image acquisition conditions. Additionally, the completeness of identified RTSs from the two proposed unsupervised methods remains lower than that achieved by fine-tuning based domain adaptive segmentation.

With the rapid advancement of domain adaptation methods in computer science, future research can explore more sophisticated model architectures for mapping multi-temporal RTSs. For example, incorporating optimal transport concepts may enhance performance by finding transformations that minimize domain discrepancies (e.g., Ackaouy et al., 2020; Courty et al., 2016). Additionally, models designed for complex target domains, for example, multi-target domains (Gholami et al., 2020) and dynamic target domains (Wu & He, 2022), can be adapted for multi-temporal RTS mapping. Moreover, large visual foundation models like the Segment Anything Model, trained on an unprecedented scale of data sets, exhibit high generalization capabilities and may effectively address domain shift issues in multi-temporal RTS mapping (Kirillov et al., 2023).

In addition to modifying model structures, automated post-processing based on multi-source data can help filter out false positives. Altitude changes caused by RTSs can be detected through Interferometric Synthetic Aperture Radar or high-resolution Digital Elevation Model (DEM) data sets, allowing for the exclusion of false positives that lack significant change (Bernhard et al., 2020; Huang et al., 2021). Furthermore, certain feature extraction methods can filter out false positives that do not exhibit the typical geometry of RTSs (Huang et al., 2023).

## 7. Conclusion

Remote sensing images provide invaluable information for RTS mapping in large-scale, inaccessible environment. Yet visual interpretation is labor intensive, and classic semantic segmentation models heavily rely on ground truth. In addition, their performance reduces drastically when the location and time are shifted. Aiming at improving the model transfer generalization at a low cost, we first demonstrate and quantitatively assess the performance degradation of model trained on source domain. Then we incorporated domain adaptation into deep learning model. In the supervised domain adaptation, we applied the classical fine-tuning approach; whereas for the UDA, we constructed the ISTDA Framework and the TADA network.

Our results demonstrate that: (a) The proposed three domain adaptation methods can improve the temporal generalization ability in given target domains; (b) TADA excels the others in terms of zero label cost and high

accuracy in complex transfer scenarios; (c) ISTDA does not present satisfactory performance when the target domain deviates from the source domain; (d) Fine-tuning based domain adaptive segmentation, at the cost of manual data annotation, can adequately map slumps in complex transfer scenarios, and it surpasses the two unsupervised methods when the IoU threshold is high.

Our research offers an UDA method (TADA), achieving remarkable temporal generalization high at no cost of labels. Our research proves the immense potential of domain adaptation methods in monitoring large-scale, multi-temporal geomorphological features.

## Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

## Data Availability Statement

The PlanetScope images used as input data of deep learning model are Data for this research are not publicly available due to the commercial use restrictions. The images are stored in this in-text data citation reference: Planet Team (2017). The Copernicus Global DEM data used for analyzing topographical influences are available in the in-text data citation reference: European Space Agency, Sinergise (2021). The environmental data sets include soil moisture and soil types are available in these in-text data citation references: Liu and Zhang (2022) and Shangguan et al. (2023) after registration. We have archived preserved our RTS ground truth data and codes for deep learning model construction, pre-processing, post-processing and figure generation in Zenodo (Lin et al., 2024).

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 42371078), Distinguished Young Scholars of the National Natural Science Foundation of China (Overseas), Emerging Engineering Interdisciplinary Youth Program at Peking University, the Fundamental Research Funds for the Central Universities, Program of the State Key Laboratory of Cryospheric Science and Frozen Soil Engineering, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences (no. CSFSE-KF-2414), the Second Tibetan Plateau Scientific Expedition and Research Program (no. 20190ZKK0905), and CUHK Direct Grant for Research (no. 4052335).

## References

- Ackaouy, A., Courty, N., Vallée, E., Commowick, O., Barillot, C., & Galassi, F. (2020). Unsupervised domain adaptation with optimal transport in Multi-Site segmentation of multiple sclerosis lesions from MRI data. *Frontiers in Computational Neuroscience*, 14. <https://doi.org/10.3389/ficom.2020.00019>
- Bernhard, P., Zwieback, S., & Hajnsek, I. (2022). Accelerated mobilization of organic carbon from retrogressive thaw slumps on the northern Taymyr Peninsula. *The Cryosphere*, 16(7), 2819–2835. <https://doi.org/10.5194/tc-16-2819-2022>
- Bernhard, P., Zwieback, S., Leinss, S., & Hajnsek, I. (2020). Mapping retrogressive thaw slumps using single-pass TanDEM-X observations. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 3263–3280. <https://doi.org/10.1109/jstars.2020.3000648>
- Burn, C. R., & Lewkowicz, A. G. (1990). Canadian landform examples- 17 retrogressive thaw slumps. *Canadian Geographer/Le Géographe Canadien*, 34(3), 273–276. <https://doi.org/10.1111/j.1541-0064.1990.tb01092.x>
- Chadha, A., & Andreopoulos, Y. (2019). Improved techniques for adversarial discriminative domain adaptation. *IEEE Transactions on Image Processing*, 29, 2622–2637. <https://doi.org/10.1109/tip.2019.2950768>
- Chen, L., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with atrous separable convolution for semantic image segmentation. In *Lecture notes in computer science* (pp. 833–851). [https://doi.org/10.1007/978-3-030-01234-2\\_49](https://doi.org/10.1007/978-3-030-01234-2_49)
- Chen, N., Tsedbazar, N., Hamunyela, E., Verbesselt, J., & Herold, M. (2021). Sub-annual tropical forest disturbance monitoring using harmonized Landsat and Sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation*, 102, 102386. <https://doi.org/10.1016/j.jag.2021.102386>
- Cheng, G., Li, S., Nan, Z., & Tong, B. (2011). Map of permafrost on the Qinghai-Tibet Plateau (1:3,000,000) (1983–1996) [Dataset]. *National Tibetan Plateau Third Pole Environment Data Center*. <https://doi.org/10.11888/Geocry.tpdcc.270014>
- Cheng, G., & Wu, T. (2007). Responses of permafrost to climate change and their environmental significance, Qinghai-Tibet Plateau. *Journal of Geophysical Research*, 112(F2). <https://doi.org/10.1029/2006jf000631>
- Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9), 1853–1865. <https://doi.org/10.1109/tipami.2016.2615921>
- European Space Agency, Sinergise. (2021). Copernicus global digital elevation model [Dataset]. *OpenTopography*. <https://doi.org/10.5069/G9028PQB>
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*. <https://doi.org/10.5555/294645.2946704>
- Gholami, B., Sahu, P., Rudovic, O., Bousmalis, K., & Pavlovic, V. (2020). Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29, 3993–4002. <https://doi.org/10.1109/tip.2019.2963389>
- Gilbert, G., Kanevskiy, M., & Murton, J. B. (2016). Recent advances (2008–2015) in the study of ground ice and cryostratigraphy. *Permafrost and Periglacial Processes*, 27(4), 377–389. <https://doi.org/10.1002/ppp.1912>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., D, W.-F., Ozair, S., & Bengio, Y. (2016). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
- Hayes, S., Lim, M., Whalen, D., Mann, P., Fraser, P., Penlington, R., & Martin, J. E. (2022). The role of massive ice and exposed headwall properties on retrogressive thaw slump activity. *Journal of Geophysical Research: Earth Surface*, 127(11). <https://doi.org/10.1029/2022jf006602>
- Hjort, J., Streletskiy, D. A., Doré, G., Wu, Q., Bjella, K., & Luoto, M. (2022). Impacts of permafrost degradation on infrastructure. *Nature Reviews Earth & Environment*, 3(1), 24–38. <https://doi.org/10.1038/s43017-021-00247-8>

- Huang, L., Liu, L., Luo, J., Lin, Z., & Niu, F. (2021). Automatically quantifying evolution of retrogressive thaw slumps in Beiluhe (Tibetan Plateau) from multi-temporal CubeSat images. *International Journal of Applied Earth Observation and Geoinformation*, 102, 102399. <https://doi.org/10.1016/j.jag.2021.102399>
- Huang, L., Luo, J., Lin, Z., Niu, F., & Liu, L. (2020). Using deep learning to map retrogressive thaw slumps in the Beiluhe region (Tibetan Plateau) from CubeSat images. *Remote Sensing of Environment*, 237, 111534. <https://doi.org/10.1016/j.rse.2019.111534>
- Huang, L., Willis, M. J., Li, G. N., Lantz, T. C., Schaefer, K., Wig, E., et al. (2023). Identifying active retrogressive thaw slumps from ArcticDEM. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205, 301–316. <https://doi.org/10.1016/j.isprsjprs.2023.10.008>
- Islam, K. A., Uddin, M. S., Kwan, C., & Li, J. (2020). Flood detection using multi-modal and multi-temporal images: A comparative study. *Remote Sensing*, 12(15), 2455. <https://doi.org/10.3390/rs12152455>
- Jorgenson, M. T. (2013). *Treatise on geomorphology: Thermokarst terrains* (pp. 313–324). Elsevier Science.
- Kirillov, A. M., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. Paper presented at Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France. <https://doi.org/10.1109/iccv51070.2023.00371>
- Kokelj, S. V., Kokoszka, J., Van Der Sluijs, J., Rudy, A., Tunnicliffe, J., Shakil, S., et al. (2021). Thaw-driven mass wasting couples slopes with downstream systems, and effects propagate through Arctic drainage networks. *The Cryosphere*, 15(7), 3059–3081. <https://doi.org/10.5194/tc-15-3059-2021>
- Kuang, X., & Jiao, J. J. (2016). Review on climate change on the Tibetan Plateau during the last half century. *Journal of Geophysical Research: Atmospheres*, 121(8), 3979–4007. <https://doi.org/10.1002/2015jd024728>
- Lewkowicz, A. G., & Harris, C. (2005). Frequency and magnitude of active-layer detachment failures in discontinuous and continuous permafrost, northern Canada. *Permafrost and Periglacial Processes*, 16(1), 115–130. <https://doi.org/10.1002/ppp.522>
- Lewkowicz, A. G., & Way, R. G. (2019). Extremes of summer climate trigger thousands of thermokarst landslides in a high Arctic environment. *Nature Communications*, 10(1), 1329. <https://doi.org/10.1038/s41467-019-109314-7>
- Li, R., Wu, Q., Li, X., Sheng, Y., Hu, G., Cheng, G., et al. (2019). Characteristic, changes and impacts of permafrost on Qinghai-Tibet Plateau. *Chinese Science Bulletin*, 64(27), 2783–2795. <https://doi.org/10.1360/tb-2019-0191>
- Li, X., Du, Z., Huang, Y., & Tan, Z. (2021). A deep translation (GAN) based change detection network for optical and SAR remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 179, 14–34. <https://doi.org/10.1016/j.isprsjprs.2021.07.007>
- Liebel, L., & Körner, M. (2018). Auxiliary tasks in multi-task learning. *arXiv Preprint. arXiv:1805.06334*.
- Lin, Y., Hu, X., Lu, H., Niu, F., Liu, G., Huang, L., et al. (2024). YilingLin0610/Multi-temporal-RTS-mapping: RTS (RTS). Zenodo. <https://doi.org/10.5281/zenodo.12892619>
- Lin, Z., Burn, C. R., Niu, F., Luo, J., Liu, M., & Yin, G. (2015). The thermal regime, including a reversed thermal offset of arid permafrost sites with variations in vegetation cover density, Wudaoliang Basin, Qinghai-Tibet Plateau. *Permafrost and Periglacial Processes*, 26(2), 142–159. <https://doi.org/10.1002/ppp.1840>
- Lin, Z., Gao, Z., Fan, X., Niu, F., Luo, J., Yin, G., & Liu, M. (2020). Factors controlling near surface ground-ice characteristics in a region of warm permafrost, Beiluhe Basin, Qinghai-Tibet Plateau. *Geoderma*, 376, 114540. <https://doi.org/10.1016/j.geoderma.2020.114540>
- Liu, F., & Zhang, G. (2022). Dataset of digital soil mapping products for the Qinghai-Tibet Plateau (2015–2024) [Dataset]. *National Tibetan Plateau Third Pole Environment Data Center*. <https://cstr.cn/18406-11.Terre.tpdc.272482>
- Liu, Y., & Li, X. (2014). Domain adaptation for land use classification: A spatio-temporal knowledge reusing method. *ISPRS Journal of Photogrammetry and Remote Sensing*, 98, 133–144. <https://doi.org/10.1016/j.isprsjprs.2014.09.013>
- Lu, P., Han, J., Yi, Y., Tu, H., Zhou, F., Meng, X., et al. (2023). MT-INSAR unveils dynamic permafrost disturbances in Hoh Xil (Kekexili) on the Tibetan Plateau hinterland. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–16. <https://doi.org/10.1109/tgrs.2023.3253937>
- Luo, J., Niu, F., Lin, Z., Liu, M., & Yin, G. (2015). Thermokarst lake changes between 1969 and 2010 in the Beilu river basin, Qinghai-Tibet Plateau, China. *Science Bulletin*, 60(5), 556–564. <https://doi.org/10.1007/s11434-015-0730-2>
- Luo, J., Niu, F., Lin, Z., Liu, M., & Yin, G. (2019). Recent acceleration of thaw slumping in permafrost terrain of Qinghai-Tibet Plateau: An example from the Beiluhe Region. *Geomorphology*, 341, 79–85. <https://doi.org/10.1016/j.geomorph.2019.05.020>
- Luo, J., Niu, F., Lin, Z., Liu, M., Yin, G., & Gao, Z. (2022). Inventory and frequency of retrogressive thaw slumps in permafrost region of the Qinghai-Tibet Plateau. *Geophysical Research Letters*, 49(23). <https://doi.org/10.1029/2022gl099829>
- Ma, Y., Chen, S., Ermon, S., & Lobell, D. B. (2024). Transfer learning in environmental remote sensing. *Remote Sensing of Environment*, 301, 113924. <https://doi.org/10.1016/j.rse.2023.113924>
- Matasci, G., Longbotham, N., Pacifici, F., Kanevski, M., & Tuia, D. (2015). Understanding angular effects in VHR imagery and their significance for urban land-cover model portability: A study of two multi-angle in-track image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107, 99–111. <https://doi.org/10.1016/j.isprsjprs.2015.05.004>
- Mu, C., Abbott, B. W., Norris, A. J., Mu, M., Fan, C., Chen, X., et al. (2020). The status and stability of permafrost carbon on the Tibetan Plateau. *Earth-Science Reviews*, 211, 103433. <https://doi.org/10.1016/j.earscirev.2020.103433>
- Nitze, I., Heidler, K., Barth, S., & Grosse, G. (2021). Developing and testing a deep learning approach for mapping retrogressive thaw slumps. *Remote Sensing*, 13(21), 4294. <https://doi.org/10.3390/rs13214294>
- O'Neill, H. B., & Burn, C. R. (2012). Physical and temporal factors controlling the development of near-surface ground ice at Illisarvik, western Arctic coast, Canada. *Canadian Journal of Earth Sciences*, 49(9), 1096–1110. <https://doi.org/10.1139/e2012-043>
- Othman, E., Bazi, Y., Melgani, F., Alhichri, H., Alajlan, N., & Zuaier, M. (2017). Domain adaptation network for cross-scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(8), 4441–4456. <https://doi.org/10.1109/tgrs.2017.2692281>
- Pan, L., Xia, H., Yang, J., Niu, W., Wang, R., Song, H., et al. (2021). Mapping cropping intensity in Huaihe basin using phenology algorithm, all Sentinel-2 and Landsat images in Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, 102, 102376. <https://doi.org/10.1016/j.jag.2021.102376>
- Pan, S. J., Tsang, I. W., Kwok, J. T., & Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2), 199–210. <https://doi.org/10.1109/tnn.2010.2091281>
- Peng, J., Huang, Y., Sun, W., Chen, N., Ning, Y., & Du, Q. (2022). Domain adaptation in remote sensing image classification: A survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 9842–9859. <https://doi.org/10.1109/jstars.2022.3220875>
- Planet Team. (2017). Planet application program interface: In space for life on Earth [Dataset]. <https://api.planet.com>
- Ramage, J., Irrgang, A., Herzschuh, U., Morgenstern, A., Couture, N., & Lantuit, H. (2017). 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. <https://doi.org/10.1002/2017jf004231>
- Ran, Y., Cheng, G., Dong, Y., Hjort, J., Lovecraft, A. L., Kang, S., et al. (2022). Permafrost degradation increases risk and large future costs of infrastructure on the Third Pole. *Communications Earth & Environment*, 3(1), 238. <https://doi.org/10.1038/s43247-022-00568-6>

- Rodenizer, H., Yang, Y., Fiske, G., Potter, S., Windholz, T., Mullen, A., et al. (2024). A comparison of satellite imagery sources for automated detection of retrogressive thaw slumps. *Remote Sensing*, 16(13), 2361. <https://doi.org/10.3390/rs16132361>
- Rudy, A., Lamoureux, S. F., Kokelj, S. V., Smith, I. R., & England, J. (2017). Accelerating thermokarst transforms Ice-Cored terrain triggering a downstream cascade to the ocean. *Geophysical Research Letters*, 44(21). <https://doi.org/10.1002/2017gl074912>
- Runge, A., Nitze, I., & Grosse, G. (2022). Remote sensing annual dynamics of rapid permafrost thaw disturbances with LandTrendr. *Remote Sensing of Environment*, 268, 112752. <https://doi.org/10.1016/j.rse.2021.112752>
- Shangguan, Y., Shi, Z., & Min, X. (2023). A 1 km daily soil moisture dataset over the Qinghai-Tibet Plateau (2001–2020) [Dataset]. *National Tibetan Plateau Third Pole Environment Data Center*. <https://cstr.cn/18406.11.Terre.tpdc.300224>
- Su, Y., Zhang, G., Mei, S., Lian, J., Wang, Y., & Wan, S. (2023). Reconstruction-assisted and distance-optimized adversarial training: A defense framework for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–13. <https://doi.org/10.1109/tgrs.2023.3328889>
- Tong, B., Li, S., Bo, J., & Qiu, G. (2011). Permafrost map along at the 1:600 000 in the Tibet highway [Dataset]. *National Tibetan Plateau Third Pole Environment Data Center*. <https://doi.org/10.11888/Geocry.tpdc.270621>
- Tuia, D., Persello, C., & Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2), 41–57. <https://doi.org/10.1109/mgrs.2016.2548504>
- Turetsky, M. R., Abbott, B. W., Jones, M. C., Anthony, K. M. W., Olefeldt, D., Schuur, E. A. G., et al. (2020). Carbon release through abrupt permafrost thaw. *Nature Geoscience*, 13(2), 138–143. <https://doi.org/10.1038/s41561-019-0526-0>
- Turner, K. W., Pearce, M. D., & Hughes, D. D. (2021). 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. <https://doi.org/10.3390/rs13020171>
- Walter, K. M., Edwards, M. E., Grosse, G., Zimov, S. A., & Chapin, F. S., III. (2007). Thermokarst lakes as a source of atmospheric CH<sub>4</sub> during the last deglaciation. *Science*, 318(5850), 633–636. <https://doi.org/10.1126/science.1142924>
- Wang, J., Ma, A., Zhong, Y., Zheng, Z., & Zhang, L. (2022). Cross-sensor domain adaptation for high spatial resolution urban land-cover mapping: From airborne to spaceborne imagery. *Remote Sensing of Environment*, 277, 113058. <https://doi.org/10.1016/j.rse.2022.113058>
- Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, 312, 135–153. <https://doi.org/10.1016/j.neucom.2018.05.083>
- Wang, S., Guan, K., Zhang, C., Jiang, C., Zhou, Q., Li, K., et al. (2023). Airborne hyperspectral imaging of cover crops through radiative transfer process-guided machine learning. *Remote Sensing of Environment*, 285, 113386. <https://doi.org/10.1016/j.rse.2022.113386>
- Wang, Y., Feng, L., Zhang, Z., & Tian, F. (2023). An unsupervised domain adaptation deep learning method for spatial and temporal transferable crop type mapping using Sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 199, 102–117. <https://doi.org/10.1016/j.isprsjprs.2023.04.002>
- Witharana, C., Udawalpol, M. R., Liljedahl, A. K., Jones, M. K. W., Jones, B. M., Hasan, A., et al. (2022). Automated detection of retrogressive thaw slumps in the High Arctic using high-resolution satellite imagery. *Remote Sensing*, 14(17), 4132. <https://doi.org/10.3390/rs14174132>
- Wu, J., & He, J. (2022). Domain adaptation with dynamic open-set targets. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 2039–2049). <https://doi.org/10.1145/3534678.3539235>
- Wu, J., Lin, L., Li, T., Cheng, Q., Zhang, C., & Shen, H. (2022). Fusing Landsat 8 and Sentinel-2 data for 10-m dense time-series imagery using a degradation-term constrained deep network. *International Journal of Applied Earth Observation and Geoinformation*, 108, 102738. <https://doi.org/10.1016/j.jag.2022.102738>
- Xia, Z., Huang, L., Fan, C., Jia, S., Lin, Z., Liu, L., et al. (2022). Retrogressive thaw slumps along the Qinghai-Tibet engineering corridor: A comprehensive inventory and their distribution characteristics. *Earth System Science Data*, 14(9), 3875–3887. <https://doi.org/10.5194/essd-14-3875-2022>
- Yang, D., Qiu, H., Ye, B., Liu, Y., Zhang, J., & Zhu, Y. (2023). Distribution and recurrence of warming-induced retrogressive thaw slumps on the central Qinghai-Tibet Plateau. *Journal of Geophysical Research: Earth Surface*, 128(8). <https://doi.org/10.1029/2022jf007047>
- Yang, Y., Rogers, B. M., Fiske, G., Watts, J. D., Potter, S., Windholz, T., et al. (2023). Mapping retrogressive thaw slumps using deep neural networks. *Remote Sensing of Environment*, 288, 113495. <https://doi.org/10.1016/j.rse.2023.113495>
- Yu, W., Han, F., Liu, W., & Harris, S. A. (2016). Geohazards and thermal regime analysis of oil pipeline along the Qinghai-Tibet Plateau Engineering Corridor. *Natural Hazards*, 83(1), 193–209. <https://doi.org/10.1007/s11069-016-2308-y>
- Zhang, T., Zhang, X., Zhu, P., Jia, X., Tang, X., & Jiao, L. (2022). Generalized few-shot object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 353–364. <https://doi.org/10.1016/j.isprsjprs.2022.12.004>
- Zhao, Z., Tang, P., Zhao, L., & Zhang, Z. (2021). Few-Shot object detection of remote sensing images via two-stage fine-tuning. *IEEE Geoscience and Remote Sensing Letters*, 19, 1–5. <https://doi.org/10.1109/lgrs.2021.3116858>
- Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232. <https://doi.org/10.1109/iccv.2017.244>
- Zwieback, S., Kokelj, S. V., Günther, F., Boike, J., Grosse, G., & Hajnsek, I. (2018). Sub-seasonal thaw slump mass wasting is not consistently energy limited at the landscape scale. *The Cryosphere*, 12(2), 549–564. <https://doi.org/10.5194/tc-12-549-2018>