CROSS-MODAL CHANGE DETECTION FLOOD EXTRACTION BASED ON SELF-SUPERVISED CONTRASTIVE PRE-TRAINING

Wenqing Feng[1,*], Fangli Guan[1], Chenhao Sun[2], Wei Xu[3]

[1] School of Computer Science, Hangzhou Dianzi University, Hangzhou, P.R. China
(Corresponding author, e-mail: wq_feng@whu.edu.cn)
[2] Electrical & Information Engineering School, Changsha University of Science & Technology, Changsha, P.R. China
[3] Information System and Management College, National University of Defense Technology, Changsha, P.R. China

ISPRS TC I Mid-term Symposium

KEY WORDS: Cross-modal, Change detection, Flood extraction, Self-supervised contrastive learning, Pre-training

ABSTRACT:

Flood extraction is a critical issue in remote sensing analysis. Accurate flood extraction faces challenges such as complex scenes, image differences across modalities, and a shortage of labeled samples. Traditional supervised deep learning algorithms demonstrate promising prospects in flood extraction. They mostly rely on abundant labeled data. However, in practical applications, there is a scarcity of available labeled samples for flood change regions, leading to an expensive acquisition of such data for flood extraction. In contrast, there is a wealth of unlabeled data in remote sensing images. Self-supervised contrastive learning (SSCL) provides a solution, allowing learning from unlabeled data without explicit labels. Inspired by SSCL, we utilized the open-source CAU-Flood dataset and developed a framework for cross-modal change detection in flood extraction (CMCDFE). We employed the Barlow Twin (BT) SSCL algorithm to learn effective visual feature representations of flood change regions from unlabeled cross-modal bi-temporal remote sensing data. Subsequently, these well-initialized weight parameters were transferred to the task of flood extraction, achieving optimal accuracy. We introduced the improved CS-DeepLabV3+ network for extracting flood change regions from cross-modal bi-temporal remote sensing data, incorporating the CBAM dual attention mechanism. By demonstrating on the CAU-Flood dataset, we proved that fine-tuning with only a pre-trained encoder can surpass widely used ImageNet pre-training methods without additional data. This approach effectively addresses downstream cross-modal change detection flood extraction tasks.

1. INTRODUCTION

In recent years, frequent global flood disasters have caused substantial damage to both property and community safety. The essence of flood extraction lies in recognizing floodwaters, specifically in determining the extent of inundation (Zhang et al., 2021). With the advancement of satellite remote sensing technology, remote sensing images have become crucial tools, providing essential means and data support for acquiring flood-related information. Their rapid acquisition, strong timeliness, and capacity for large-scale repetitive observations significantly contribute to flood monitoring. Both multispectral remote sensing images and synthetic aperture radar (SAR) remote sensing images are applied in flood monitoring. The fusion of multispectral remote sensing images and SAR remote sensing images harnesses the complementary advantages of each, thereby enhancing the effectiveness of remote sensing-based flood monitoring (He et al., 2023; Zhang et al., 2021; Zhao et al., 2023; Zhang et al., 2023).

Since the introduction of fully convolutional networks (Long et al., 2014) in 2015, a multitude of end-to-end deep learning methodologies has been integrated into the task of cross-modal change detection (CD) for flood extraction, playing a pivotal role. These deep learning approaches primarily rely on supervised learning, demanding a substantial volume of labeled data (Konapala et al., 2021; Zhao et al., 2023; Zhang et al., 2023). However, flood disasters are characterized by their sudden and transient nature, with a scarcity of high-resolution satellite imagery during such events. Additionally, annotating remote sensing images incurs a high cost, making the acquisition of well-labeled flood samples a time-consuming and labor-intensive endeavor. To mitigate reliance on annotated data, various cross-modal flood extraction methods utilize pre-processing models on large-scale ImageNet datasets, followed by fine-tuning with a limited amount of pixel-level annotations. However, substantial distribution disparities between ImageNet data and cross-modal flood monitoring datasets pose a considerable risk of domain shift issues. Recently, self-supervised learning has garnered significant research interest in academia as a method to derive effective visual representations from a vast pool of unlabeled images (Caron et al. 2020 ; Chen et al., 2020; Chen and He 2021; Grill et al. 2020; He et al. 2019; Tian et al. 2019; Zbontar et al. 2021). Essentially, self-supervised learning consists of two steps: firstly, it involves a pretext task, where well-designed self-supervised signals and pseudo-labels (i.e., automatically generated labels) are utilized to aid in initializing model parameters. This enables the model to directly extract rich visual feature knowledge from unlabeled image data. Subsequently, this acquired knowledge is transferred to specific downstream tasks to reduce reliance on a large number of labeled samples, thereby enhancing the model's performance in those tasks.

Self-supervised contrastive pre-training represents a novel feature learning paradigm, primarily focused on defining positive and negative sample pairs. Its main objective is to maximize the similarity within positive pairs while minimizing it for negative pairs in the feature space, embodying the principle of "attraction within the same class, exclusion between different classes." Recent research highlights the high generalization ability of features pre-trained by most self-supervised contrastive learning (SSCL) methods, a notable advantage in initializing backbone networks for downstream

https://doi.org/10.5194/isprs-annals-X-1-2024-75-2024 | © Author(s) 2024. CC BY 4.0 License.

tasks. Among the early SSCL methods, MoCo (He et al. 2019) and SimCLR (Chen et al., 2020) introduced innovative concepts like the momentum encoder and queue sampling, effectively handling negative samples. However, MoCo's computational speed is somewhat hindered due to queue sampling, while SimCLR, with its larger batch size, may incur higher GPU memory demands and increased computational costs to prevent the learning of trivial solutions. BYOL (Grill et al. 2020) addresses trivial solutions through positive sample contrastive learning, employing a symmetric network and stop-gradient methods. Nonetheless, it faces challenges related to sensitivity to task requirements and hyper-parameter tuning. SimSiam (Chen and He 2021), which avoids negative samples, leverages an asymmetric network structure and cross-gradient updates to counteract trivial solutions. However, it requires additional computational resources and exhibits sensitivity to hyper-parameters. SwAV (Caron et al. 2020), rooted in online clustering and multi-view prediction encoding, successfully circumvents negative sample requirements, proving advantageous in high-label-cost scenarios. Nevertheless, it does entail higher computational resources due to the intricate nature of the online clustering algorithm. In contrast, Barlow Twin (BT) (Zbontar et al. 2021) introduces an innovative approach to SSCL, unencumbered by batch size restrictions and negative samples. It emphasizes the embedding itself, steering clear of asymmetric structure design. By computing the cross-correlation matrix of augmented samples and utilizing a loss function to mitigate redundancy, BT achieves a cross-correlation matrix reminiscent of an identity matrix. This indicates that the feature vectors of different augmentations of the same sample exhibit similarity, thereby minimizing redundancy across different dimensions and enhancing feature representation efficiency.

Given the inherent strengths of BT, we have chosen to employ it as the objective function for SSCL within our proposed framework for cross-modal CD flood extraction (CMCDFE). This correspondence presents the results of experimental validation conducted using the publicly available CAU-Flood dataset (He et al., 2023). CAU-Flood stands out as a remote sensing dataset explicitly crafted for cross-modal flood extraction, featuring multiple sets of pre-disaster Sentinel-2 optical images, post-disaster Sentinel-1 SAR images, and corresponding ground truth label images delineating altered regions. The articulated CMCDFE framework unfolds in two sequential phases: self-supervised contrastive pre-training and fine-tuning. In the initial stage, we construct a three-channel false-color image by amalgamating post-disaster Sentinel-1 VV polarization mode data, near-infrared band images extracted from pre-disaster Sentinel-2 optical images, and computed NDWI (Normalized Difference Water Index) index images. Subsequently, the BT algorithm is enlisted to distill effective visual representations of altered areas from these unlabeled false-color images. In the subsequent stage, we leverage the SSCL methodology to pre-train the encoder of the refined CS-DeepLabV3+ model. The encoder demonstrates noteworthy parameter initialization, and empirical evidence derived from the CAU-Flood flood monitoring dataset attests that fine-tuning exclusively with the pre-trained encoder outperforms the widely embraced ImageNet pre-training approach, eliminating the need for additional data. This methodological refinement efficiently addresses downstream tasks associated with cross-modal flood extraction.

The remainder of this paper is structured as follows. Section 2 outlines the proposed methodology, providing a detailed description. In Section 3, we present the results of our experiments and engage in a comprehensive discussion. The concluding remarks are offered in Section 4 to wrap up this paper.

2. METHODOLOGY

2.1 CMCDFE framework

We utilized pre-disaster Sentinel-2 multispectral images to extract near-infrared band images and selected NDWI as the representation of water bodies in the multispectral data. NDWI is computed by calculating data from the near-infrared and green bands, and its calculation formula is as follows:

$$\mathrm{NDWI} = \frac{\mathrm{Green} - \mathrm{NIR}}{\mathrm{Green} + \mathrm{NIR}} \tag{1}$$

This paper synthesizes post-disaster Sentinel-1 VV polarization mode data, near-infrared band images, and NDWI index images through channel fusion, constructing a three-channel false-color image. Assuming $\mathbf{X}_{\text{train}}$ is the synthesized three-channel false-color image, we use the BT SSCL algorithm to pre-train the model on an unlabeled training set $\left\{\mathrm{U} = \left(\mathbf{X}_{\text{train}}\right)_i\right\}_{i=1}^{N}$. This enables the pre-trained encoder to easily fine-tune on the labeled training set $\left\{\mathrm{L} = \left(\mathbf{X}_{\text{train}}, \mathbf{Y}_{\text{train}}\right)_i\right\}_{i=1}^{N}$. Our goal is to use self-supervised contrastive pre-training to learn effective visual feature representations of cross-modal flood changes from the synthesized three-channel false-color images. Subsequently, the learned encoder weights are used as the initial weights for the downstream flood change region extraction task network, the CS-DeepLabV3+ algorithm. Figure 1 illustrates the schematic diagram of the proposed CMCDFE framework, where the knowledge transfer of self-supervised contrastive learning feature representation is well-validated in the downstream task.

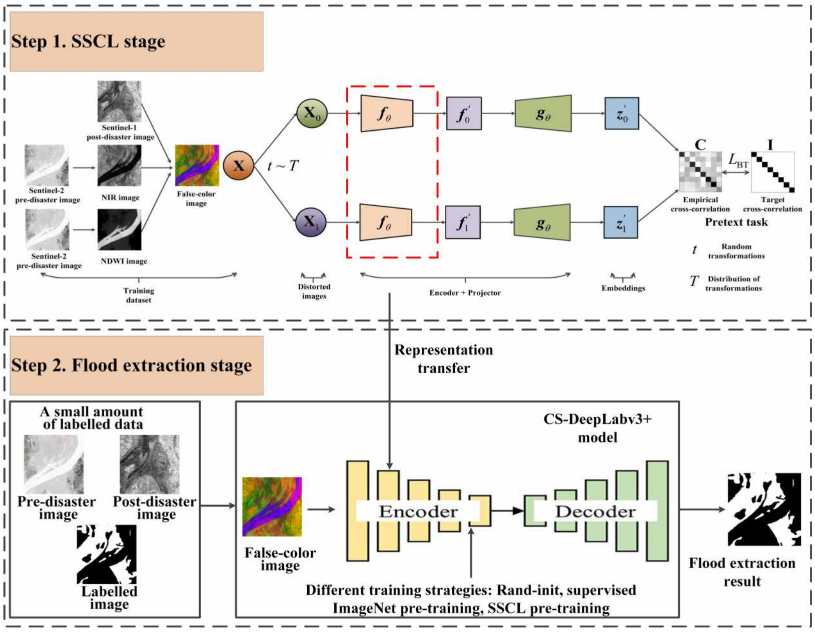

Figure 1. Pipeline of the proposed CMCDFE framework.

2.2 SSCL pre-training method

The method employed in this paper can be divided into two parts. Firstly, the BT algorithm is utilized for SSCL pre-training. Subsequently, the pre-trained weights are transferred to downstream cross-modal flood extraction. The BT algorithm maintains basic consistency with the SimCLR model in several aspects, including image augmentation, the encoder, and the projection module. It employs ResNet50 (excluding the final classification layer) as the feature extractor, followed by a projector network. This projector network consists of two linear layers, each with a hidden layer size of 512 output units. Due to high computational requirements, the output of the projection network is modified to generate embeddings of size 256, whereas the original BT network produces embeddings of size 8192 (Zbontar et al. 2021). The first layer of the projector is followed by a batch normalization layer and rectified linear units. Figure 2 provides an overview of the BT algorithm. As the BT algorithm does not require distinguishing between positive and negative samples, the input false-color image $\mathbf{X}$ undergoes transformations $t \sim T$ to obtain different augmented data $\mathbf{X}_0$ and $\mathbf{X}_1$. After passing through the encoder $\boldsymbol{f}_\theta$, they respectively yield features $\boldsymbol{f}_0^{'}$ and $\boldsymbol{f}_1^{'}$. Following the projector layer $\boldsymbol{g}_\theta$, the extracted features are denoted as $\boldsymbol{z}_0^{'}$ and $\boldsymbol{z}_1^{'}$. For a given batch, the network's loss function is:

$$L_{\mathrm{BT}} \triangleq \underbrace{\sum_i \left(1-C_{ii}\right)^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j\neq i} C_{ij}^2}_{\text{redundancy reduction term}} \tag{2}$$

The initial component of the loss function is denoted as the "invariance term," while the subsequent one is termed the "redundancy reduction term." Here, C signifies the cross-correlation matrix and can be computed as follows:

$$C_{ij} \triangleq \frac{\sum_b \left(Z_1^{'}\right)_{b,i} \left(Z_2^{'}\right)_{b,j}}{\sqrt{\sum_b \left(\left(Z_1^{'}\right)_{b,i}\right)^2} \sqrt{\sum_b \left(\left(Z_2^{'}\right)_{b,j}\right)^2}} \tag{3}$$

Where C_{ii} denotes the diagonal elements of the cross-correlation matrix C, C_{ij} represents the non-diagonal elements, and λ is a hyperparameter. Here, the parameter b represents different batch samples, indicating that the calculation of each element in C is conducted across the batch dimensions. It can be observed that the optimization objective aims for the diagonal elements of the cross-correlation matrix C to be 1, and the non-diagonal elements to be 0. After multiple training iterations, the cross-correlation matrices calculated for positive examples of the same image under different transformations tend to approach the identity matrix (Zbontar et al. 2021).

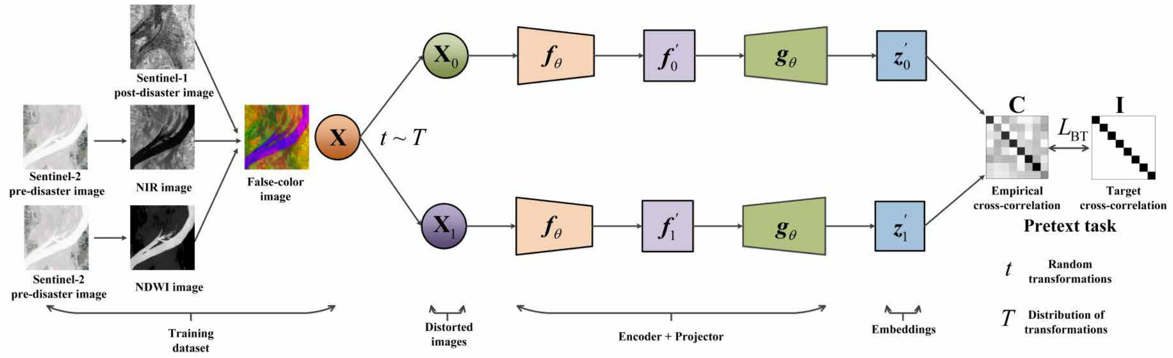

Figure 2. Flow chart of the proposed SSCL pre-training algorithm.

2.3 CS-DeepLabV3+ algorithm

DeepLabV3+ (Chen et al., 2018) is a semantic segmentation model based on convolutional neural networks, designed to address challenges in detail capture and multi-scale feature extraction for image segmentation tasks. It incorporates techniques such as dilated convolutions and atrous spatial pyramid pooling (ASPP). This paper introduces the CBAM (Convolutional Block Attention Module) attention mechanism (Woo et al., 2018) into the enhanced DeepLabV3+ model, referred to as the CS-DeepLabV3+ model, as depicted in Figure 3. The model aims to better focus on crucial information related to cross-modal flood extraction tasks, thereby improving the network's precision in locating flood change areas. The CS-DeepLabV3+ network structure comprises two main components: the encoder and the decoder. In the encoder stage, false-color images undergo down-sampling through ResNet 50 for the main feature extraction network, completing the capture of image features. The high-dimensional feature maps generated by the main feature network are then input into the ASPP module. The processed outputs are overlaid, concatenated, and channel reduction is achieved through 1×1 convolutions to reduce computational complexity. Finally, the CBAM attention mechanism is applied for weighted operations, enabling high-level feature extraction and multi-scale information integration. In the decoder stage, the input image undergoes a 4x down-sampling to capture low-order features with rich details. Channel adjustments are made using 1×1 convolutions, and the CBAM attention mechanism is applied to process low-level features, filtering out background information and highlighting flood change areas. Subsequently, the high-level features obtained in the encoding phase undergo a 4x up-sampling to match the size of low-level feature maps. The low and high-level features are then concatenated, followed by channel adjustment through 3×3 convolutions to achieve feature fusion. Finally, a 4x up-sampling is performed to restore spatial information, generating flood change area extraction result maps consistent with the input image size.

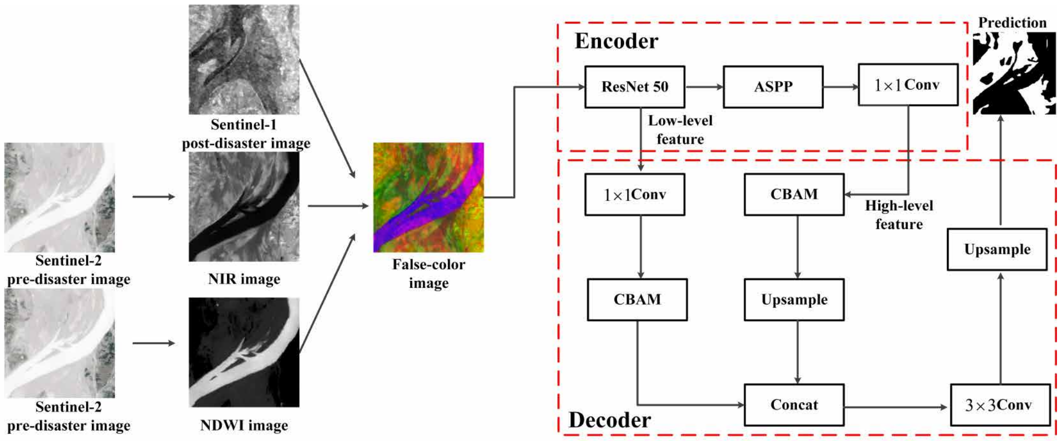

Figure 3. Structure of the proposed CS-DeepLabV3+ algorithm.

3. EXPERIMENTAL ANALYSES AND DISCUSSION

3.1 Dataset description and evaluation metrics

To evaluate the efficacy of the proposed algorithm, we employed the CAU-Flood cross-modal flood extraction dataset, comprising pre-disaster Sentinel-2 optical images and post-disaster Sentinel-1 SAR images for 18 distinct study regions. Encompassing a comprehensive area of 95,142 square kilometers, the CAU-Flood dataset spans diverse geographical locations, including China, Bangladesh, Australia, the United States, Canada, and Germany. Notably, the Sentinel-1 images exhibit spatial resolutions ranging from 3.5 meters to 40 meters. Radar images from Sentinel-1 were acquired at the Ground Range Detected (GRD) level, with exclusive processing of data from the VV-polarization mode to optimize flood detection accuracy. Sentinel-2 images consist of four bands (red, green, blue, and near-infrared) with a spatial resolution of 10 meters. Ensuring semantic consistency in cross-modal interpretation, the CAU-Flood dataset underwent resampling to enforce uniform image sizes for SAR and optical pairs, and grayscale values were stretched to a standardized range of 0 to 255 (He et al., 2023). Processed Sentinel-2 and Sentinel-1 images served as pre-disaster and post-disaster inputs, respectively, with manual annotations identifying flood areas. This yielded a dataset comprising 18,302 image patches sized at 256×256, with 15,231 patches allocated for training and 3,071 for testing.

During the pre-training phase, we utilized pairs of pre-disaster and post-disaster images from the training set to construct false-color images, employing the BT algorithm for SSCL without the use of labeled images. Subsequently, the pre-trained ResNet50 encoder segment, endowed with well-tuned parameters, was transferred to the downstream CS-DeepLabV3+ model for the cross-modal flood extraction task. Four metrics, namely precision, recall, F1 score, and IoU, were employed to evaluate our algorithm's performance, and comparisons were made with state-of-the-art (SOTA) methods.

3.2 Implementation details

This paper utilizes the PyTorch framework to implement the BT algorithm. The specifications of our experimental machine are as follows: a 12th Gen Intel Core i9-12900K @ 3.19 GHz processor, 64.00 GB RAM, and an NVIDIA GeForce RTX 3090 graphics card. To optimize the model, we adhere to the BT protocol (Zbontar et al. 2021). During the SSCL pre-training phase, we set the batch size to 20 and employed the LARS optimizer for training over 400 epochs. The initial learning rate was set at 0.005, adjusted by multiplying with the batch size and dividing by 256. We introduced a learning rate warm-up period of 10 epochs, followed by a cosine decay schedule (Zbontar et al. 2021), reducing the learning rate by a factor of 1000. The trade-off parameter of the loss function is set to $\lambda = 5\times10^{-3}$, and the weight decay parameter value is 1.5×10^{-6}.

https://doi.org/10.5194/isprs-annals-X-1-2024-75-2024 | © Author(s) 2024. CC BY 4.0 License.

We evaluated the performance of the CS-DeepLabV3+ algorithm on downstream cross-modal flood extraction using the SSCL and fine-tuning strategies. For consistency, we employed ResNet50 as the feature extractor for the CS-DeepLabV3+ algorithm. In the downstream task experiments of this paper, we chose to use the Adam optimizer with beta1 set to 0.9, beta2 set to 0.999, and epsilon set to 1.0×10^{-8}. The initial learning rate was set to 0.005, with a batch size of 10, and a total of 150 epochs were conducted. We adopted a hybrid loss function (Fang et al., 2022), combining weighted cross-entropy and Dice loss with equal weights.

3.3 Results and discussion

3.3.1 Evaluation on CAU-Flood dataset

We conducted a comparison between the cross-modal flood extraction method CS-DeepLabV3+ proposed in this paper and several SOTA methods, including UNet++ (Peng et al., 2019), ResUNet (Zhang et al., 2018), PSPNet (Zhao et al., 2017), HRNet (Sun et al., 2019), and DeeplabV3+. Among them, both UNet++ and ResUNet represent advancements over the traditional UNet network, which has become a fundamental network in various remote sensing applications. UNet++ inherits the structure of UNet while incorporating dense skip connections, thereby maximizing the preservation of fine-grained details and global information. In contrast, ResUNet leverages the benefits of both residual networks and UNet, with residual connections alleviating the gradient vanishing problem in deep networks, contributing to faster convergence and improved training efficiency. PSPNet aggregates contextual information from different regions of the image using pyramid pooling modules, integrating complex contextual information into the pixel-level semantic segmentation framework. HRNet transforms the connection between high-resolution and low-resolution feature maps from a serial to a parallel structure, thereby maintaining the representation of high-resolution feature maps throughout the entire network. To assess the effectiveness of the SSCL pre-training method proposed in this paper, we initialized the parameters of five comparative methods during training using ImageNet pre-trained weights. In contrast, our approach involves transferring the pre-trained weights of the BT algorithm to the cross-modal flood extraction task.

The results of cross-modal flood extraction on the CAU-Flood dataset using various SOTA methods are presented in Figure 4. From top to bottom, these scenarios include pre-disaster Sentinel-2 multispectral imagery, post-disaster Sentinel-1 VV polarization mode data, false-color images, ground truth images (where white represents changed areas and black represents unchanged areas), and the results of UNet++, ResUNet, PSPNet, HRNet, DeeplabV3+, and CS-DeepLabV3+. The results in Figure 4 demonstrate that all six comparative methods are proficient in handling the cross-modal flood extraction task, with each SOTA method yielding satisfactory segmentation results. Nonetheless, there are instances of suboptimal segmentation outcomes. While each model exhibits some missed detections, the extracted results overall remain acceptable. In Figure 4, grey represents true-negative (TN) pixels, green represents true-positive (TP) pixels, blue indicates false-positive (FP) pixels, and red corresponds to false-negative (FN) pixels. The qualitative comparison results in Figure 4 demonstrate that flood detection based on deep learning exhibits good adaptability to different types of land cover and can be employed in situations with frequent flooding. It can effectively identify flooded areas in environments such as estuaries, inland river plains, villages, and lakes. The flood detection accuracy evaluation results of these comparative methods on the test set are presented in Table 1. Thanks to the CAU-Flood dataset, various deep learning models demonstrate excellent performance in the cross-modal flood extraction task. The proposed CS-DeepLabV3+ method achieves the best results (Precision = 0.9315, Recall = 0.9388, F1 = 0.9351, IoU = 0.8781), as evident from Table 1. CS-DeepLabV3+ generates finer contours that are more consistent with the ground truth. Both quantitative and qualitative comparative analyses further support the superiority of the proposed method in this study.

Methods	CAU-Flood			
	Precision	Recall	F1	IoU
UNet++	**93.37**	91.61	92.48	86.02
ResUNet	91.74	**94.07**	92.89	86.72
PSPNet	91.25	93.85	92.53	86.10
HRNet	92.41	93.65	93.02	86.96
DeeplabV3+	92.95	93.23	93.09	87.07
Proposed (CS-DeepLabV3+ with BT)	93.15	93.88	**93.51**	**87.81**

Table 1. Performance comparison for the CAU-Flood dataset. (All values are in percentages.)

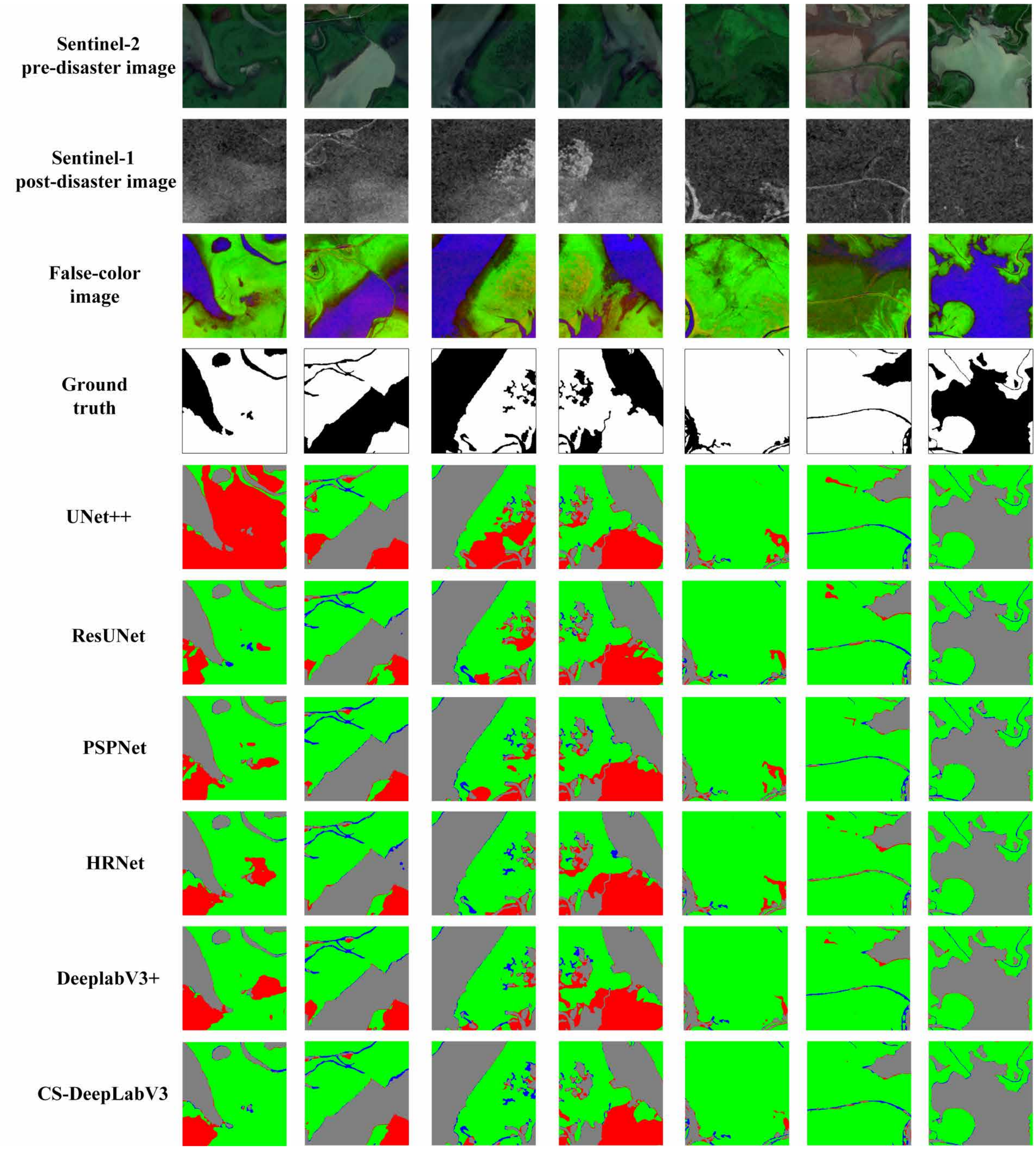

Figure 4. Visual comparisons of the different SOTA models applied to CAU-Flood. Grey: TN pixels; green: TP pixels; blue: FP pixels; red: FN pixels.

3.3.2 Ablation experiment

We fine-tuned the proposed CS-DeepLabV3+ algorithm on the CAU-Flood dataset using three strategies: random initialization (Rand-init), supervised ImageNet pre-training (ImageNet-sup), and pre-training with BT self-supervised learning. To assess the effectiveness of the adopted self-supervised contrastive pre-training method, we conducted a detailed comparative analysis with three additional self-supervised pre-training methods, namely SimCLR (Chen et al., 2020), MoCo (He et al., 2019), and CMC (Tian et al. 2019). These methods all employ a contrastive loss, which differs from the BT loss function used in our approach. According to the results in Table 2, our proposed BT self-supervised contrastive pre-training method outperforms the widely used ImageNet pre-training method, as well as SimCLR, MoCo, and CMC methods. Compared to the Rand-init strategy, on the CAU-Flood test set, applying our BT pre-training process increased the F1 score by nearly 1.5%. Similarly, our proposed pre-training method slightly outperformed the supervised ImageNet pre-training method and other SSCL methods in terms of performance. Overall, ablation experiments indicate that our proposed method for unlabeled cross-modal remote sensing images can achieve or even surpass the performance of widely used ImageNet pre-training methods and SSCL methods such as SimCLR, MoCo, and CMC, which utilize over a million labeled images. These results also indirectly confirm that our method mitigates the domain shift problem caused by transfer learning from ImageNet weights in the task of cross-modal flood extraction.

Methods		Pre-training	ImageNet dataset	CAU-Flood			
				Precision	Recall	F1	IoU
CS-DeepLabV3+ (Backbone ResNet50)	Rand-init	✕	✕	90.98	93.20	92.08	85.32
	ImageNet-sup	✓	✓	**93.68**	92.92	93.30	87.44
	SimCLR	✓	✓	93.27	93.53	93.40	87.62
	MoCo	✓	✓	93.17	93.75	93.46	87.72
	CMC	✓	✓	93.26	93.05	93.15	87.18
	BT	✓	✕	93.15	**93.88**	**93.51**	**87.81**

Table 2. Ablation experiment results on CAU-Flood dataset. (All values are in percentages. ✕ indicates excluded steps during the training process, while ✓ denotes their inclusion.)

3.3.3 Efficiency under limited labels

As is well known, large-scale flood extraction tasks currently face a challenge due to the absence of a sufficiently large and publicly available annotated dataset. This limitation hinders the widespread application of deep learning methods in cross-modal flood extraction tasks. On one hand, annotating cross-modal bi-temporal remote sensing images for flood change regions from large-scale datasets is an expensive, tedious, time-consuming, and primarily manual process. On the other hand, there is an urgent need for methods capable of learning and expressing visual information in cross-modal images without the reliance on labeled samples. To thoroughly validate the performance of the proposed BT self-supervised contrastive pre-training method with a small number of labeled samples, we specifically fine-tune a cross-modal flood extraction network with a limited number of labeled samples and evaluate the accuracy of the final flood extraction results. This paper compares the impact of Rand-init, ImageNet-sup, SimCLR, MoCo, and CMC on the performance of cross-modal flood extraction tasks, as presented in Table 3. From Table 3, it can be observed that in the CAU-Flood dataset used in the experiment, the BT self-supervised contrastive pre-training method consistently outperforms other compared pre-training methods in terms of Recall, F1, and IoU values. It is noteworthy that by using only 5% of the training labeled samples from the CAU-Flood dataset, corresponding Precision, Recall, F1, and IoU values of 87.59%, 91.87%, 89.68%, and 81.29%, respectively, can be achieved. These results indirectly demonstrate the effectiveness of our self-supervised contrastive pre-training method in addressing the problem of insufficient labeled data. Furthermore, the experimental results further confirm that employing the BT self-supervised contrastive pre-training method enables the learning of additional discriminative feature information from unlabeled image samples in the research area, which is highly beneficial for downstream cross-modal flood extraction tasks. Transferring the learned optimal parameters to the improved CS-DeepLabV3+ network significantly enhances the performance of downstream tasks.

Methods	5% of the Labeled Samples			
	CAU-Flood			
	Precision	Recall	F1	IoU
Rand-init	86.78	91.38	89.02	80.21
ImageNet-sup	87.53	90.95	89.21	80.52
SimCLR	**87.60**	90.93	89.23	80.56
MoCo	87.47	91.37	89.38	80.80
CMC	86.90	91.30	89.05	80.26
BT	87.59	**91.87**	**89.68**	**81.29**

Table 3. Performance of the different pre-training methods evaluated using the improved CS-DeepLabV3+ model with limited labels. (All values are in percentages.)

4. CONCLUSION

Supervised deep learning models demand a substantial amount of annotated data when tasked with flood extraction from cross-modal remote sensing images. However, the collection and annotation of samples containing the desired flood change regions are both time-consuming and labor-intensive. To tackle this challenge, the adoption of transfer learning with a self-supervised contrastive pre-training strategy has proven effective. In this study, we applied the BT self-supervised learning algorithm to learn effective visual feature representations of flood change regions from unlabeled cross-modal bi-temporal remote sensing data. Subsequently, these well-initialized weight parameters were transferred to the task of flood extraction. We introduced an improved CS-DeepLabV3+ network for extracting flood change regions from cross-modal bi-temporal remote sensing data, incorporating the CBAM dual attention mechanism. Experimental analysis on the open-source CAU-Flood dataset validated the effectiveness of our proposed method. The results demonstrated that fine-tuning with only a pre-trained encoder can surpass widely used ImageNet pre-training methods without the need for additional data, effectively addressing downstream cross-modal flood extraction tasks. Even with a limited number of labeled data samples, our self-supervised pre-training strategy proves effective. This proves particularly beneficial for flood extraction applications facing challenges in acquiring labeled data for flood change regions due to cost constraints. In the future, we plan to replace the ResNet50 encoder component of our approach with a vision transformer to further enhance the accuracy of flood extraction.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant Nos. 42101358.

REFERENCES

Chen, T., S. Kornblith, M. Norouzi, and G. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. arXiv:2002.05709.

Chen, X., and He, K.M. 2021. Exploring Simple Siamese Representation Learning. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15745–15753).

Caron, M., I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *Advances in Neural Information Processing Systems*, 33: 9912–9924.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018;pp. 801–818.

Fang, S., K. Li, J. Shao, and Z. Li. 2022. SNUNet-CD: A Densely Connected Siamese Network for Change Detection of VHR Images. IEEE Geoscience & Remote Sensing Letters 19:1–5.

Grill, J., F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch. et al. 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *Advances in neural information processing systems*, 33:21271–21284.

He, X.N., Zhang, S.C., Xue, B.W., Zhao, T., Wu, T. 2023. Cross-modal change detection flood extraction based on convolutional neural network. *International Journal of Applied Earth Observation and Geoinformation*, 117, 103197.

He, K., H. Fan, Y. Wu, S. Xie, and R. Girshick. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729–9738).

Konapala, G., Kumar, S.V., Ahmad, S.K., 2021. Exploring Sentinel-1 and Sentinel-2 diversity for flood inundation mapping using deep learning. *ISPRS J. Photogramm. Remote Sens*, 180, 163–173.

Long, J., E. Shelhamer, and T. Darrell. 2014. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (4): 640–651.

Peng, D., Zhang, Y., Guan, H. 2019. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sensing*, 11, 1382.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Wang, J. 2019. High-resolution representations for labeling pixels and regions. arXiv 2019, arXiv:1904.04514.

Tian, Y., Krishnan, D., Isola, P. Contrastive Multiview Coding. arXiv 2019, arXiv:1906.05849.

Woo, S., Park, J., Lee, J. Y., Kweon, I. S. 2018. CBAM: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV) (pp. 3-19).

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J. 2017. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.

Zhang, Z., Liu, Q., 2018. Wang, Y. Road extraction by deep residual u-net. *IEEE Geosci. Remote Sens. Lett.*, 15, 749–753.

Zbontar, J., L. Jing, I. Misra, Y. LeCun, and S. Deny. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In International Conference on Machine Learning (pp. 12310–12320). PMLR.

Zhang, L.C., Xia, J.S. 2021. Flood detection using multiple Chinese satellite datasets during 2020 China summer floods. *Remote Sensing*, 14(1), 51.

Zhao, B.F., Sui, H.G., Liu, J.Y. 2023. Siam-DWENet: Flood inundation detection for SAR imagery using a cross-task transfer siamese network. *International Journal of Applied Earth Observation and Geoinformation*, 116, 103132.

Zhang, Y.W., Liu, P., Chen, L.J., Xu, M.Z., Guo, X.Y., Zhao, L.J. 2023. A new multi-source remote sensing image sample dataset with high resolution for flood area extraction: GF-FloodNet, *International Journal of Digital Earth*, 16:1, 2522-2554.