






Article

An Improved Method for Human Activity Detection with High-Resolution Images by Fusing Pooling Enhancement and Multi-Task Learning

Haoji Li ¹, Shilong Ren ^{1,*} , Lei Fang ¹ , Jinyue Chen ¹ , Xinfeng Wang ¹ , Guoqiang Wang ^{1,2} ,
Qingzhu Zhang ¹ and Qiao Wang ¹

¹ Academician Workstation for Big Data in Ecology and Environment, Environment Research Institute, Shandong University, Qingdao 266003, China

² Center for Geodata and Analysis, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China

* Correspondence: slren@sdu.edu.cn

Abstract: Deep learning has garnered increasing attention in human activity detection due to its advantages, such as not relying on expert knowledge and automatic feature extraction. However, the existing deep learning-based approaches are primarily confined to recognizing specific types of human activities, which hinders scientific decision-making and comprehensive environmental protection. Therefore, there is an urgent need to develop a deep learning model to address multiple-type human activity detection with finer-resolution images. In this study, we proposed a new multi-task learning model (named PE-MLNet) to simultaneously achieve change detection and land use classification in GF-6 bitemporal images. Meanwhile, we also designed a pooling enhancement module (PEM) to accurately capture multi-scale change details from the bitemporal feature maps through combining differencing and concatenating branches. An independent annotated dataset at Yellow River Delta was taken to examine the effectiveness of PE-MLNet. The results showed that PE-MLNet exhibited obvious improvements in both detection accuracy and detail handling compared with other existing methods. Further analysis uncovered that the areas of buildings, roads, and oil depots has obviously increased, while the farmland and wetland areas largely decreased over the five years, indicating an expansion of human activities and their increased impacts on natural environments.

Keywords: human activity; change detection; semantic segmentation; multi-task learning; remote sensing



Academic Editors: Tengfei Long, Wei Jiang, Xing Wang, Elhadi Adam and Jungho Im

Received: 19 December 2024

Revised: 3 January 2025

Accepted: 4 January 2025

Published: 5 January 2025

Citation: Li, H.; Ren, S.; Fang, L.; Chen, J.; Wang, X.; Wang, G.; Zhang, Q.; Wang, Q. An Improved Method for Human Activity Detection with High-Resolution Images by Fusing Pooling Enhancement and Multi-Task Learning. *Remote Sens.* **2025**, *17*, 159. <https://doi.org/10.3390/rs17010159>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Human activities encompass a diverse range of undertakings conducted continuously for survival, development, and the enhancement of living standards [1]. With the acceleration of industrialization and urbanization globally, human activities have increasingly impacted the natural environment, leading to environmental pollution and ecological degradation, such as deforestation, lake reclamation [2], and overexploitation [3]. Consequently, real-time monitoring of human activities is crucial for urban planning and sustainable development.

Two primary means of monitoring human activities are aerial remote sensing and satellite remote sensing [4]. Aerial remote sensing, leveraging aircraft or other airborne platforms equipped with multi-band scanners, cameras, and radars, offers high accuracy

and flexibility but is limited in scope for large-area, long-term monitoring due to aircraft performance and economic constraints [5]. In contrast, satellite remote sensing boasts global coverage, continuous observation, and low data acquisition costs [6–8], making it widely applicable in urban planning, land management, resource exploration, and disaster monitoring [9–11]. Notably, the advent of high-resolution satellites has provided finer data sources for capturing intricate texture structures and change details and thus facilitating more precise land identification [12], although high-resolution images usually offer less spectral information compared to medium- and low-resolution images, like Landsat TM and MODIS [13].

Change detection, defined as “the process of identifying differences in the state of an object of phenomenon by observing it at different times” [14], is an effective means of extracting human activity information through bitemporal image change mapping [15]. Early change detection methods relied on visual interpretation, which was both knowledge-intensive and impractical for large-scale remote sensing data [16]. With advancements in remote sensing and computer technology, automated and intelligent change detection methods have emerged [17]. Algebra-based methods directly compute pixel differences and apply preset thresholds to generate change maps [18], while transformation-based methods use Principal Component Analysis (PCA), Tasseled Cap transformation, or Gram–Schmidt transformation to separate change components before applying thresholds [19–21]. Both approaches face challenges in accurately setting thresholds due to the complexity of and variability in remote sensing data [22]. Classification-based methods compare results after supervised or unsupervised classification results, reducing image information utilization [23]. Advanced models, like reflectance models and spectral mixture models, convert spectral information into physical parameters for change analysis, showing good performance in specific scenarios but limited generalizability [24].

Deep learning has garnered significant attention in recent years, particularly in the fields of image processing and remote sensing [25]. Compared to traditional change detection methods, deep learning approaches excel in their ability to directly learn change features from bitemporal or multi-temporal remote sensing images [26]. By leveraging these features to segment the images, deep learning models generate change maps with superior robustness [27]. Consequently, deep learning methods have been widely employed in human activity detection [28]. For instance, de Bem et al. [29] utilized three convolutional neural networks (SharpMask, U-Net, and ResUnet) to map deforestation in the Amazon region of Brazil from 2017 to 2019 and all outperformed traditional machine learning approaches. Murdaca et al. [30] introduced a semi-supervised deep learning framework to detect changes in open-pit mines, which presented excellent robustness through the incorporation of pseudo-labels. Meanwhile, D’Addabbo et al. [31] successfully detected newly constructed buildings in urban areas by leveraging pre-trained AlexNet to extract deep features and applying transfer learning techniques. Although these studies underscore the efficacy of deep learning in human activity detection, the majority of current research focuses exclusively on specific types of human activities, lacking a deep learning algorithm capable of addressing the detection of multiple types of human activities.

To address the limitations in the existing research, this study aims to develop a multi-class human activity detection model (PE-MLNet) based on high-resolution imagery utilizing a multi-task learning approach. This model not only performs accurate land use classification of bitemporal remote sensing images but also simultaneously identifies areas of change, thereby enabling comprehensive detection of human activities within the study region. Furthermore, we propose a pooling enhancement module (PEM) with a dual-branch structure for capturing both global and local details of changes in bitemporal images.

2. Materials and Methods

2.1. Study Region

The study area for this research is located in the Yellow River Delta of China. As shown in Figure 1, this region has a diverse array of land use types, encompassing natural landscapes such as rivers, wetlands, and lakes while simultaneously experiencing the impact of human activities, like urban expansion, land development, and pond aquaculture [32]. This study collected three GF-6 images covering the Yellow River Delta across June 2019, July 2021, and June 2023, which include one 2 m panchromatic band and four 8 m multispectral bands, namely, the red, green, blue, and near-infrared bands. To accurately capture the characteristics of complex land objects, we first fused the panchromatic and multispectral bands to produce multispectral imagery with a resolution of 2 m.

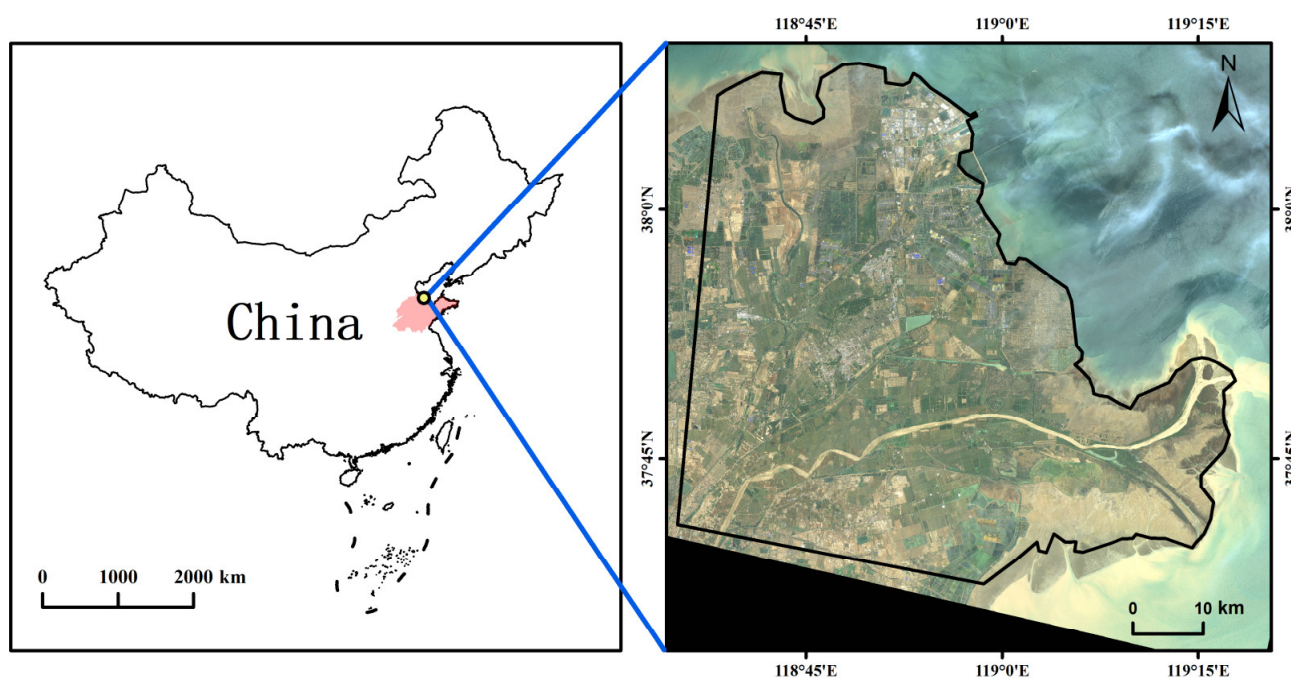


Figure 1. The location of the study area.

The sample collection strategy is illustrated in Figure 2. Specifically, 23 regions of different sizes were selected for this study (Figure 2a), covering the majority of typical human activity types and natural land covers. Manual annotation was then performed at the selected sampling points (Figure 2b). For the annotation of change detection (*Change*), a binary labeling system was adopted, with 1 representing “changed” and 0 representing “unchanged”. As for land use annotation (*Landuse1* and *Landuse2*), this study focused on nine land types related to human activities, including building, natural water, cropland, aquaculture pond, salt field, road, oil depot, wetland, and others. Subsequently, the images and samples were cropped into 256×256 tiles, resulting in 1351 sets of training data. After random partitioning, 1204 sets were designated as the training set, and 135 sets were designated as the validation set. As shown in Figure 2c, the data labels include land use labels (*label1* and *label2*) and change detection labels (*label_cd*) for bitemporal images. Table 1 provides a summary of the label categories and pixel statistics for the YRD dataset.

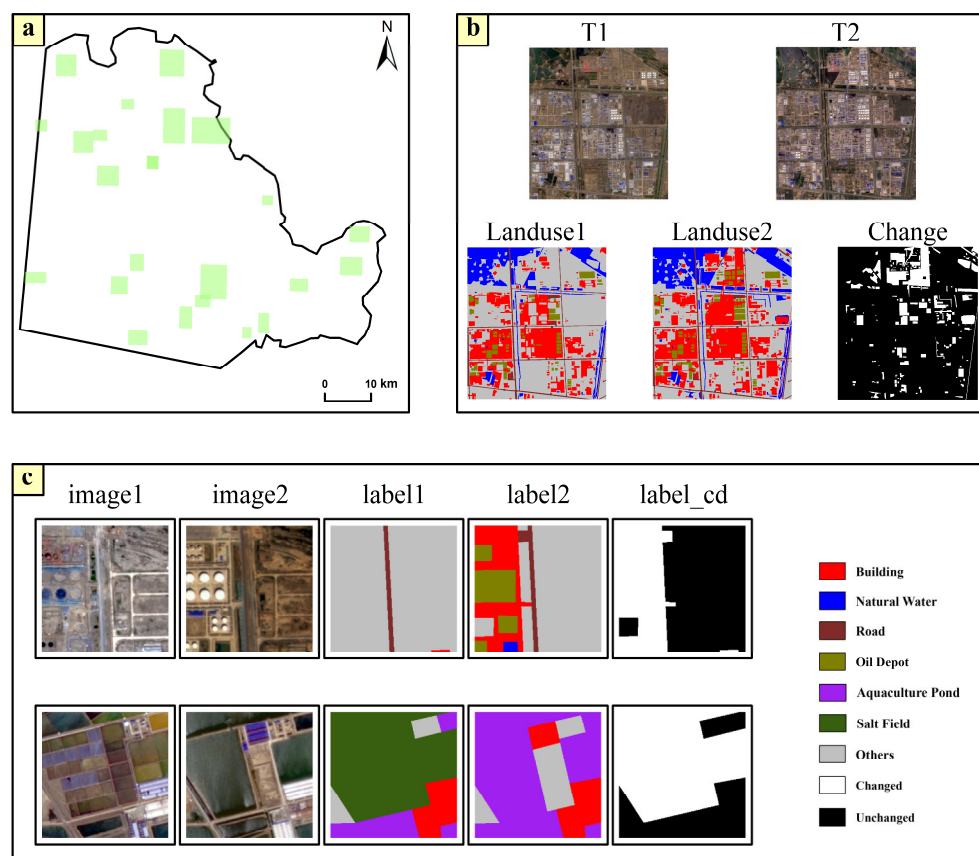


Figure 2. Sample collection strategy: (a) selected sample areas. There are a total of 23 selected green squares.; (b) detailed display of sample area annotation. The *Landuse1* and *Landuse2* are land use labels, and the *Change* is the change label. (c) Detailed display of some samples from the YRD dataset. The *label1* and *label2* are used for semantic segmentation, and the *label_cd* is used for change detection.

Table 1. The label category and the number of pixels in the YRD dataset.

Task	Class	Train	Validate
Change Detection	Changed	3.7×10^7	4.7×10^6
	Unchanged	2.0×10^8	2.2×10^7
Semantic Segmentation	Others	9.8×10^7	1.5×10^7
	Building	1.0×10^7	2.0×10^6
	Natural Water	9.4×10^7	9.4×10^6
	Cropland	1.5×10^8	1.5×10^7
	Aquaculture Pond	5.9×10^7	5.0×10^6
	Salt Field	2.3×10^7	2.0×10^7
	Road	3.1×10^6	3.5×10^5
	Oil Depot	1.3×10^6	5.0×10^5
Wetland	4.4×10^7	4.7×10^6	

2.2. Methodology

This study designed a human activity detection model called PE-MLNet based on pooling enhancement and multi-task learning. This model not only identifies the change area of dual-temporal images but also classifies the land use of the front and back images, so as to realize the detection of multiple types of human activities in the study area, with the workflow illustrated in Figure 3. It mainly includes four parts:

- (1) Feature Extractor: It extracts common semantic features from the bitemporal images.

- (2) Semantic Segmentation Module: Utilizing the feature maps obtained from the feature extractor as input, it outputs semantic segmentation results.
- (3) Pooling Enhancement Module: It captures multi-scale change details within the feature maps.
- (4) Change Detection Module: Taking the multi-scale change feature maps extracted by the PEM as input, it produces the change detection results.

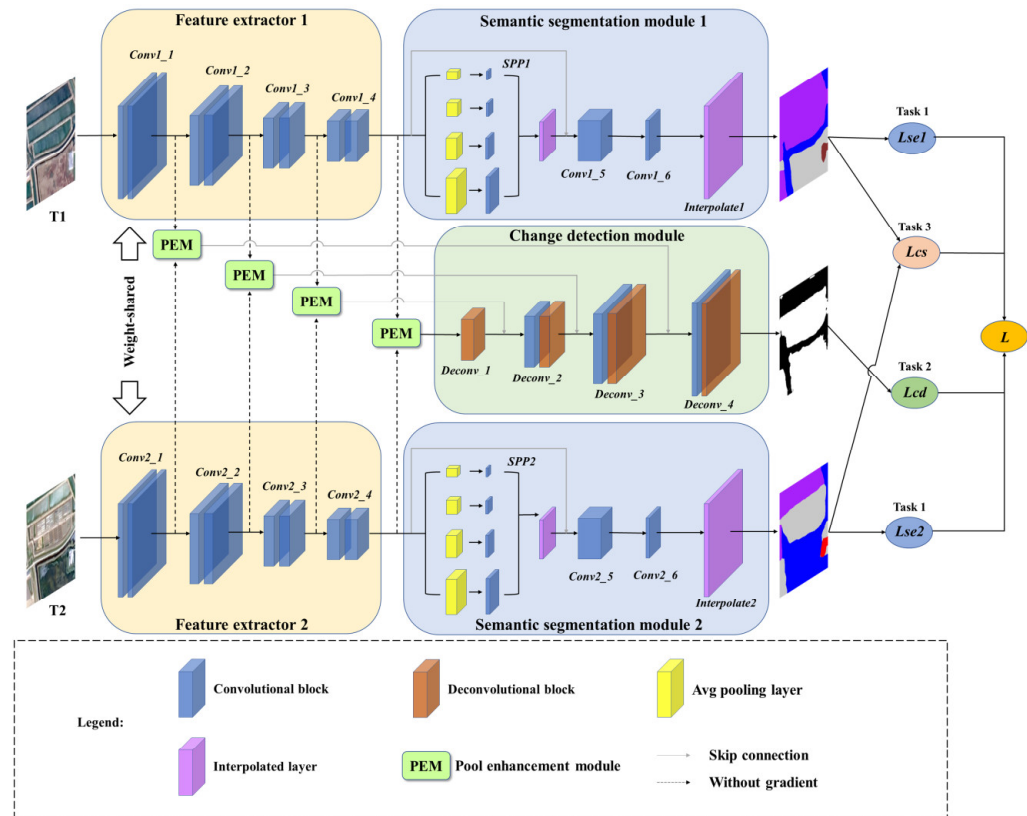


Figure 3. The structure of PE-MLNet. It consists of four parts: feature extractor, semantic segmentation module, pooling enhancement module (PEM), and change detection module.

Based on the aforementioned four modules, multi-task learning encompasses three tasks: (1) change detection in bitemporal images, (2) semantic segmentation, and (3) measuring the consistency between semantic segmentation results and change detection results.

2.2.1. Feature Extractor

In traditional change detection approaches, a single-branch network is often utilized as a feature extractor, where bitemporal images are superimposed and directly input to obtain a difference map [33]. However, this method is prone to information loss and often makes it challenging for the network to converge [34]. To address this issue, we introduced a dual-branch network based on shared parameters as the feature extractor, which is capable of better extracting semantic information and differences from bitemporal images. The branch network comprises four convolutional blocks, each consisting of Conv 3×3 , batch normalization, and ReLU. When bitemporal images are input into the feature extractor, they undergo multiple downsampling processes, resulting in a continuous reduction in the size of the feature maps and a gradual increase in the number of channels.

2.2.2. Semantic Segmentation Module

To extract semantic features from dual-phase images and obtain semantic segmentation results, the semantic segmentation module in this study adopts the structure of PSPNet [35]. The spatial pyramid pooling (SPP) module in PSPNet enhances the receptive field of the network and captures richer semantic information from the image. After the feature maps extracted by the feature extractor enter the semantic segmentation module, they first pass through the SPP module where average pooling is applied using different kernel sizes of 1×1 , 2×2 , 3×3 , and 6×6 . The channel number is then reduced using 1×1 convolutions. Subsequently, the feature maps are upsampled to restore their original dimensions and are concatenated with the feature maps before entering the SPP module. Following this, two 1×1 convolution blocks are applied to reduce the channel number without altering the dimensions and to predict the probability of each pixel belonging to different categories. Finally, the segmentation results are output after interpolation upsampling.

2.2.3. Pooling Enhancement Module (PEM)

In the field of change detection research, two common methods for generating change feature maps are differencing and concatenating. However, both approaches have inherent limitations [36]. The differencing method creates a change feature map by subtracting and then taking the absolute value. This method is susceptible to lighting and weather conditions, leading to unstable results. Additionally, the high variability in high-resolution imagery spectra introduces substantial noise during differentiation. The concatenating method connects the deep features of dual-temporal images across channels, preserving a majority of the semantic information. Yet it results in information redundancy, increases computational costs, and lacks interpretability.

Addressing the limitations of these methods, this study introduced the PEM to capture global and local change details while preserving semantic information (Figure 4). The PEM comprises concatenation and differentiation branches. Upon receiving the feature maps (F_1 and F_2) from the feature extractor, it produces a connecting feature map (A) and a differential change map (D) through respective connection and differentiation processes. The connecting feature map employs 3×3 and 1×1 convolution blocks to detect contextual information and reduce channel numbers. For the differential change map, spatial pyramid pooling is first applied, followed by average pooling using four types of receptive fields (1×1 , 2×2 , 3×3 , and 6×6). This extracts multi-scale change information from the map, which is then upsampled and combined with the original map. Finally, contextual information is detected using 3×3 convolution blocks, and pixel values are summed with those from the connecting branch. The entire process can be described as follows:

$$A = [F_1, F_2] \quad (1)$$

$$D = |F_1 - F_2| \quad (2)$$

$$A' = Conv1(Conv3(A)) \quad (3)$$

$$D' = SPP(D) \quad (4)$$

$$F' = A' + D' \quad (5)$$

where $[\cdot]$ represents concatenation on the channel dimension, $|\cdot|$ denotes the absolute value, and $Conv3$ and $Conv1$ refer to 3×3 and 1×1 convolution layers, respectively, followed by batch normalization and ReLU. PEM integrates semantic information from dual-temporal feature maps with multi-scale change details, effectively enhancing the representation of changed regions while suppressing irrelevant distractions. Through the PEM, four pooling-enhanced change maps of different scales are sequentially generated.

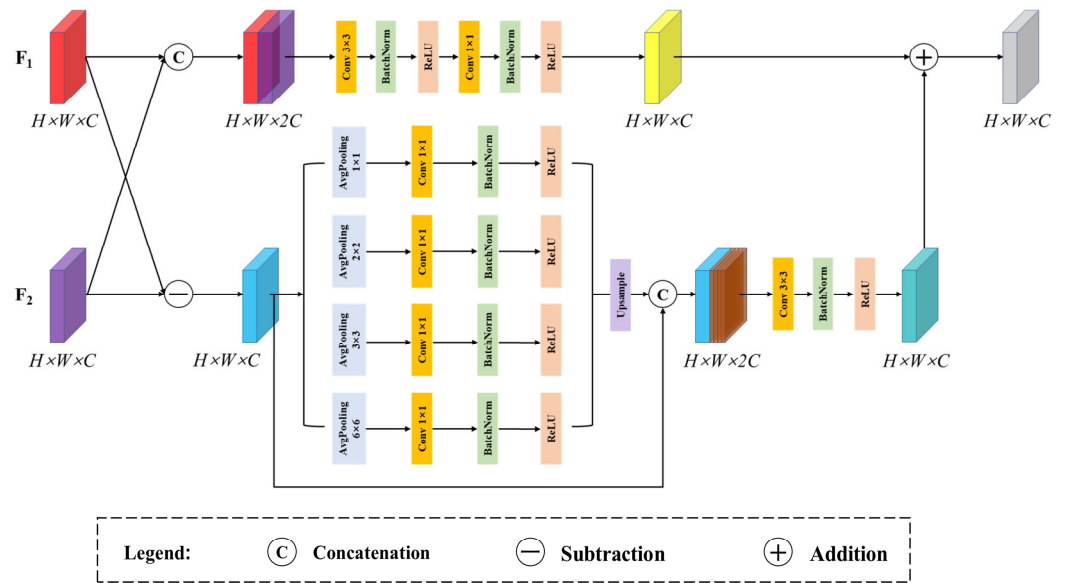


Figure 4. Illustration of pooling enhancement module.

2.2.4. Change Detection Module

The PEM extracts four types of pooling-enhanced change maps at different scales, employing skip connections to pass multiple change maps to the change detection module. The change detection module is designed as an alternating combination of convolutional and upsampling blocks. Each upsampling block comprises Transposed Conv 3×3 , batch normalization, and ReLU. After the deep change map enters the change detection module, it first reduces the channel count through convolution and then overlaps with the change map from the previous layer through upsampling. Throughout the process, four upsampling operations are performed to gradually restore the change map's size and predict the probability of pixel changes.

2.2.5. Loss Function

This study introduces a multi-task loss function to optimize multi-task learning, consisting of three components: the change detection loss L_{cd} , the semantic segmentation loss L_{se} , and the consistency loss L_{cs} between the change detection and semantic segmentation results.

The change detection loss L_{cd} quantifies the discrepancy between the binary change map I_{cd} predicted by the change detection module and $label_{cd}$. The BCELoss function is employed to compute L_{cd} , with the formula for each pixel being the following:

$$L_{cd} = -y_c \log(p_c) - (1 - y_c) \log(1 - p_c) \quad (6)$$

where y_c indicates whether the pixel has changed in $label_{cd}$, with 1 indicating change and 0 indicating no change. p_c represents the probability of the change detection module predicting a change in the pixel.

Semantic segmentation loss L_{se} represents the discrepancy between the segmentation prediction results $I1$ and $I2$ and the corresponding $label1$ and $label2$ for the dual-temporal images produced by the semantic segmentation module. In this study, L_{se} is computed using the CELoss, which is frequently employed in multi-classification tasks. The calculation formula for each pixel is as follows:

$$L_{se} = -\frac{1}{k} \sum_{i=1}^k y_i \log(p_i) \quad (7)$$

where k denotes the number of semantic segmentation categories. y_i denotes the category in *label1* or *label2*, while p_i signifies the predicted probability for category i output by the semantic segmentation module.

L_{cd} and L_{se} are, respectively, used to optimize the learning of change detection tasks and semantic segmentation tasks. Additionally, this study proposes a loss L_{cs} that connects semantic segmentation and change detection. By calculating the loss between the difference (non-zero difference results are converted to 1) of the prediction results $I1$ and $I2$ of semantic segmentation and *label_cd*, the semantic information changes in the semantic segmentation prediction results are kept consistent with the change detection prediction results. The MSE is used to calculate L_{cs} , and the calculation formula is as follows:

$$L_{cs} = \frac{1}{N} \sum_{i=1}^N (y_i - p_i)^2 \quad (8)$$

where N represents the number of pixel points, y_i is the true value of the i -th pixel point in *label_cd*, and p_i is the predicted value of the i -th pixel point in the difference result.

The loss for multi-task learning L is set as the sum of the semantic segmentation loss, change detection loss, and consistency loss.

$$L = \lambda_1 L_{cd} + \lambda_2 (L_{se_1} + L_{se_2}) + \lambda_3 L_{cs} \quad (9)$$

where L_{se_1} and L_{se_2} represent the semantic segmentation losses for each temporal phase, respectively. λ_1 , λ_2 , and λ_3 denote the hyperparameters representing weights.

3. Results and Discussion

3.1. Experimental Setup and Accuracy Assessment

The experiments in this study were conducted on a workstation equipped with an NVIDIA RTX A4000 GPU. All the models were built using the PyTorch library [37], with an iteration count of 300. The Adam algorithm [38] was chosen as the gradient optimization method, and the batch size was set to 10. To enhance the diversity and quantity of the data, random flipping and mirroring were employed. For the multi-task loss function L , considering that change detection and semantic segmentation are equally important, we hoped to achieve a balance between them during the training process. Therefore, we empirically set $\lambda_1 = 2$ and $\lambda_2 = \lambda_3 = 1$ in order to unify the magnitude of different loss values. The initial learning rate was set at 0.0001 and gradually decayed during the training process, according to the following formula:

$$lr = lr_{init} \times \left(1 - \frac{epoch}{max_epoch}\right)^{0.9} \quad (10)$$

where lr represents the current learning rate, lr_{init} denotes the initial learning rate, $epoch$ signifies the current iteration count, and max_epoch indicates the total number of iterations.

To verify the effectiveness of the model, a quantitative analysis was conducted by comparing the output results with the labels. For change detection, this study selected four evaluation metrics for analysis, i.e., precision (P), recall (R), overall accuracy (OA), and F1-score (F1). A higher P indicates fewer false positives, while a higher R signifies fewer false negatives. Both F1 and OA serve as comprehensive evaluation metrics, and higher values represent better model performance. For semantic segmentation, this study evaluated model performance using five metrics, including the Intersection over Union (IOU), mean Intersection over Union (mIOU), overall accuracy (OA) [39], Average Accuracy (AA), and Kappa coefficient (Kappa) [40]. The IOU represents the ratio of the intersection and union

between the predicted and true regions, measuring the overlap between the model's segmentation results and the ground truth labels. The mIOU is the average of the sum of the IOUs for all the categories. The AA represents the average of the sum of the recalls for each category, and the Kappa assesses the consistency of the segmentation results. The formulas are as follows:

$$P = \frac{TP}{TP + FP} \quad (11)$$

$$R = \frac{TP}{TP + FN} \quad (12)$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (14)$$

$$IOU = \frac{TP}{TP + FN + FP} \quad (15)$$

$$mIOU = \frac{1}{k} \sum_{i=1}^k IOU_i \quad (16)$$

$$AA = \frac{1}{k} \sum_{i=1}^k R_i \quad (17)$$

$$Kappa = \frac{OA - PRE}{1 - PRE} \quad (18)$$

$$PRE = \frac{(TP + FP) \times (TP + FN) \times (TN + FN) \times (TN + FP)}{(TP + TN + FP + FN)^2} \quad (19)$$

where TP , FP , TN , and FN represent the number of positive samples correctly detected, the number of negative samples detected as positive, the number of negative samples correctly detected, and the number of positive samples detected as negative, respectively. Additionally, k denotes the number of semantic segmentation categories.

3.2. Comparative Methods

To validate the effectiveness of the proposed model, this study compares it with several classic change detection networks and semantic segmentation networks. For change detection, the following networks are selected:

- (1) *Fully Convolutional-Early Fusion (FC-EF)* [41]: This approach employs an early fusion strategy, where bitemporal images are concatenated and fed into an FCN to obtain a change map. Additionally, skip connections are incorporated to supplement spatial information.
- (2) *Fully Convolutional Siamese-Concatenation (FC-Siam-conc)* [41]: Based on FC-EF, this method replaces the encoder with a dual-branch structure sharing weights. Deep features from both temporal phases are extracted, concatenated, and then input into a decoder to generate a change map.
- (3) *Fully Convolutional Siamese-Difference (FC-Siam-diff)* [41]: This structure is similar to FC-Siam-conc but with one key difference. After the encoder extracts deep features from both temporal phases, they are fused using a difference approach.
- (4) *Image Fusion Network (IFN)* [42]: This network introduces a Deep Difference Discrimination Network (DDN) based on an attention mechanism. It integrates multi-level deep features with image difference features to construct a change map.

- (5) *ChangeNet* [43]: This network utilizes ResNet to extract change information at different scales, which is then processed by a unified decoder for change detection, outputting semantically meaningful change detection results.
- (6) *DTCDSCN* [44]: This network introduces a spatial feature pyramid pooling module as its central component, which expands the receptive field of the feature maps and incorporates contextual features across different scales. Additionally, two extra semantic segmentation decoders are trained simultaneously.

For semantic segmentation, this study selects the following networks:

- (1) *Fully Convolutional Network (FCN)* [45]: Based on the VGG16 classification network, this approach replaces the fully connected layers with convolutional layers. Additionally, skip connections are added to combine deep semantic information with superficial information, aiming to generate precise segmentation results.
- (2) *UNet* [46]: This network architecture forms a symmetrical “U” shape, consisting of an encoder and a decoder. The encoder is responsible for feature extraction, while the decoder performs upsampling through deconvolutions layer by layer, restoring the feature map size and outputting the segmentation results.
- (3) *SegNet* [47]: The encoder adopts the network structure of VGG16, while the decoder utilizes pooling indices for upsampling to achieve pixel-level classification.
- (4) *HRNet* [48]: The network maintains high-resolution features of the image through the use of parallel connections while simultaneously fusing features of different resolutions through repeated information exchange modules.

3.3. Change Detection

For change detection, all the models showed a stable convergence trend during training (Figure 5a) and high accuracy over 90% (Figure 5b). This indicates that all the models are capable of effectively learning from the training data and enhancing their performance. Specifically, the FC series models exhibited a relatively fast convergence rate, stabilizing after approximately 25 epochs, with an OA reaching around 92%. The IFN model showed a higher initial loss and lower OA during the early stages of training, with significant fluctuations in OA throughout the process. Although ChangeNet experienced significant OA fluctuations in the first 50 epochs, it gradually stabilized afterward. Among all the models, PE-MLNet demonstrated the most outstanding performance. During training, the loss of PE-MLNet remained consistently lower than the other models, with an initial OA of 91% and a final OA of 97%. Additionally, the minimal fluctuation in OA further proved its excellent performance in change detection.

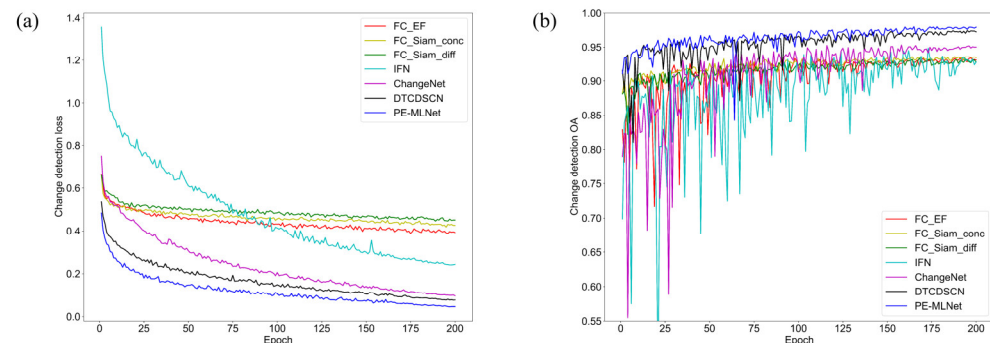


Figure 5. Modeling performance comparison. (a) Loss; (b) overall accuracy.

Furthermore, PE-MLNet also surpassed the other models in F1, precision, recall, and OA, achieving scores of 96.52%, 96.80%, 96.24%, and 98.01%, respectively (Table 2). From the visualization results (Figure 6), it can be observed that PE-MLNet is capable of extract-

ing most of the changed areas while producing fewer false positives and negatives. For instance, in Figure 6 (S1), only PE-MLNet detected the specific change from salt fields to aquaculture ponds, whereas the other methods failed to fully identify this transformation. In Figure 6 (S2 and S3), the other methods mistakenly classified unchanged roads as changed areas (S2) and failed to detect the emergence of small buildings (S3). This could be due to the narrow pixel width occupied by roads and small buildings in the imagery, leading to the loss of information during feature extraction and resulting in false detections. The PEM in PE-MLNet is designed to capture and enhance change information at different scales in the image, suppress noise, and detect subtle changes. Simultaneously, the multi-task learning framework allows the model to focus more on semantic information during feature extraction, thereby improving the model robustness.

Table 2. Quantitative comparison of change detection results among the different models. The best results are highlighted in bold.

Method	F1 (%)	Precision (%)	Recall (%)	OA (%)
FC-EF	87.68	86.17	89.41	93.29
FC-Siam-conc	88.15	86.59	89.95	93.55
FC-Siam-diff	87.00	83.84	91.38	93.28
IFN	90.13	89.72	90.91	94.85
ChangeNet	91.73	91.74	91.72	95.31
DTCDSCN	95.55	96.08	95.04	97.44
PE-MLNet	96.52	96.80	96.24	98.01

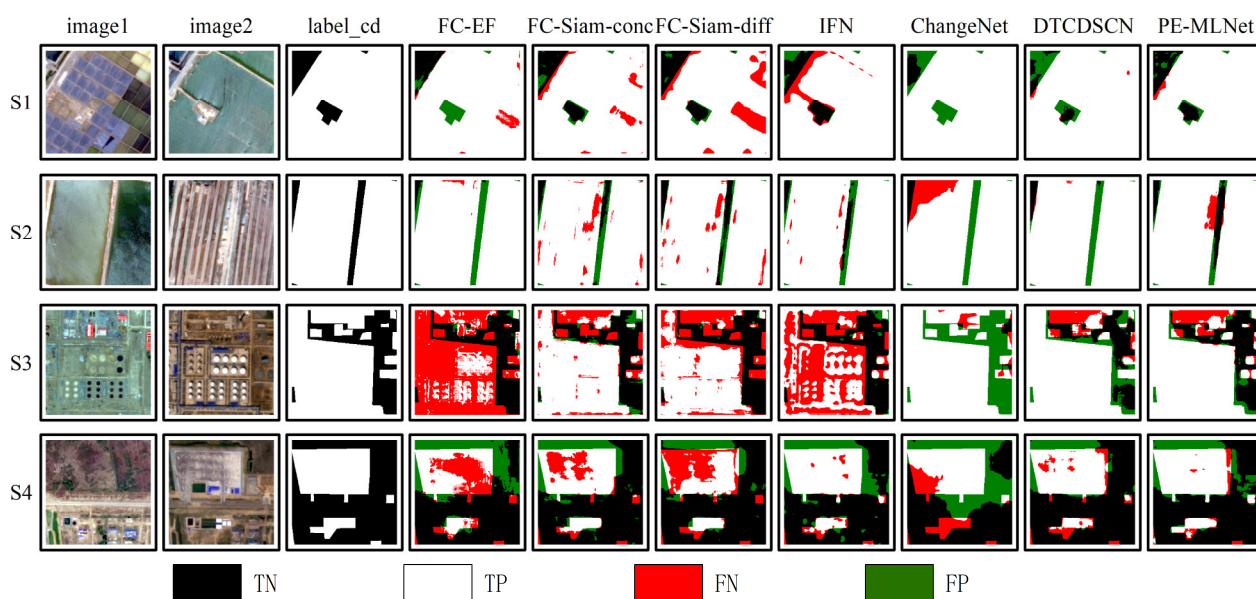


Figure 6. Visualization comparison of change detection results between PE-MLNet and other methods on the YRD dataset. S1–S4 represent four representative samples.

3.4. Semantic Segmentation

The loss of all the models rapidly decreased within the first 75 epochs and then gradually stabilized. Meanwhile, the OA of all the models steadily increased, with the highest OA reaching over 92% (Figure 7). PE-MLNet and UNet exhibited better performance than the other methods, achieving a maximum OA of over 94%. Considering more detailed evaluation metrics (Table 3), the four evaluation metrics of PE-MLNet and UNet are notably higher than those of the other methods. It is worth noting that, except for a slightly lower AA compared to UNet, PE-MLNet achieved the highest levels in the other evaluation metrics, especially in the mIOU, which is improved by 4.62%, 1.68%, 4.73%, and 1.94%

compared to the other methods, respectively. This proved the excellent performance of the model in semantic segmentation tasks.

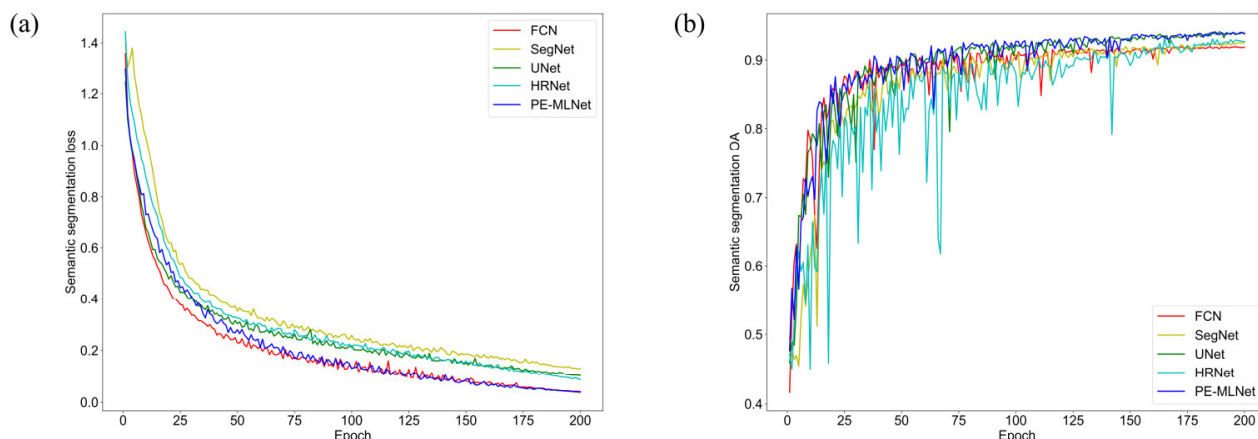


Figure 7. Modeling performance comparison for semantic segmentation. (a) Loss; (b) overall accuracy.

Table 3. Quantitative comparison of semantic segmentation results among the different models, with the best outcomes highlighted in bold.

Method	mIOU (%)	OA (%)	AA (%)	Kappa (%)
FCN	76.88	92.96	87.80	91.15
UNet	79.82	94.09	90.34	92.06
SegNet	76.77	92.94	87.43	91.17
HRNet	79.56	92.97	86.42	91.18
PE-MLNet	81.50	94.69	88.06	93.33

To compare the segmentation effects of the different models on ground object identification, the Intersection over Union (IOU) for each category was calculated (Table 4). PE-MLNet also obtained the highest IOU in multiple categories. Combined with the visualization results (Figure 8), its prediction results are more accurate in identifying various features, such as buildings, natural water bodies, and aquaculture ponds. This demonstrates a stronger robustness and stability of PE-MLNet than the others. However, for the identification of roads and oil depots, none of the models achieve satisfactory results. This may be due to the limited sample size of these two categories in the training data, making it difficult for the models to fully learn their features.

Table 4. Quantitative comparison of mean Intersection over Union (mIOU) across various categories for different models. The best results are highlighted in bold.

Method	Building	Natural Water	Cropland	Aquaculture Pond	Salt Field	Road	Oil Depot	Wetland	Others
FCN	69.47	86.05	95.37	88.95	84.62	42.46	55.23	85.87	83.88
UNet	69.88	88.63	93.19	90.22	88.06	50.14	65.39	90.62	82.25
SegNet	66.96	82.28	94.94	88.99	58.54	44.84	50.46	83.59	79.60
HRNet	70.65	83.14	95.93	93.01	83.99	49.75	63.04	87.97	89.58
PE-MLNet	72.94	89.67	96.58	94.58	89.37	50.55	63.65	89.43	86.70

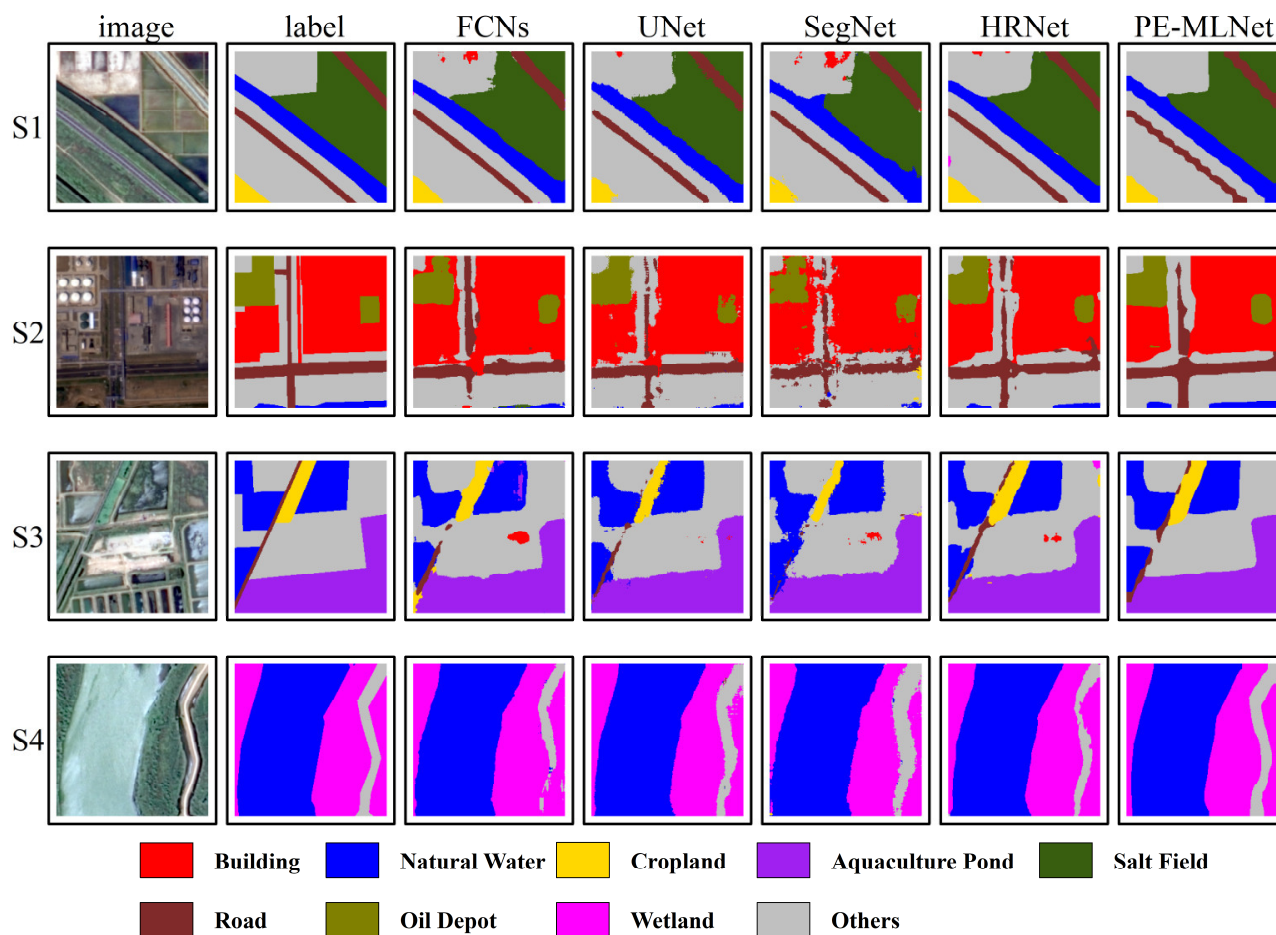


Figure 8. Visualization comparison of semantic segmentation results between PE-MLNet and other methods on the YRD dataset. S1–S4 represent four representative samples.

3.5. Effects of PEM

To evaluate the significance of the PEM in change detection, we chose to replace the PEM with two methods: concatenating (PE-MLNet_conc) and differencing (PE-MLNet_diff) of the multi-layer feature maps obtained from the feature extractor. Although the PEM increases the complexity of the model, it does not significantly impact computational efficiency. In terms of evaluation metrics, while the improvement in the PEM in OA and recall is not remarkable, it significantly outperforms the other two approaches in the F1, IOU, and precision metrics (Table 5). This suggests that the PEM plays a crucial role in optimizing the details and integrity of prediction results. It aids the network to concentrate on more detailed information during the decoding process, thereby enhancing the overall network performance. Among the three methods, PE-MLNet_diff performed worst. This could be attributed to the high variability in high-resolution spectral imagery, where relying solely on a difference approach might lead to information loss in detection results.

Table 5. Quantitative comparison of change detection results using different feature fusion methods.

Method	FLOPS (G)	Params (M)	F1 (%)	OA (%)	IOU (%)	Precision (%)	Recall (%)
PE-MLNet_PEM	37.29	34.67	96.52	98.01	93.37	96.80	96.25
PE-MLNet_conc	66.93	19.94	95.50	97.47	91.54	95.08	95.93
PE-MLNet_diff	36.21	14.81	95.39	97.39	91.36	95.31	95.48

By visualizing the Class Activation Maps (CAMs) for selected samples from the YRD dataset, as shown in Figure 9, it becomes evident that the PEM method adaptively learns weight-enhanced feature representations while accurately pinpointing the locations of changing regions. This evidence supports that the introduction of the PEM enhances the network ability to capture multi-scale detailed information, ultimately improving the precision and completeness of prediction results.

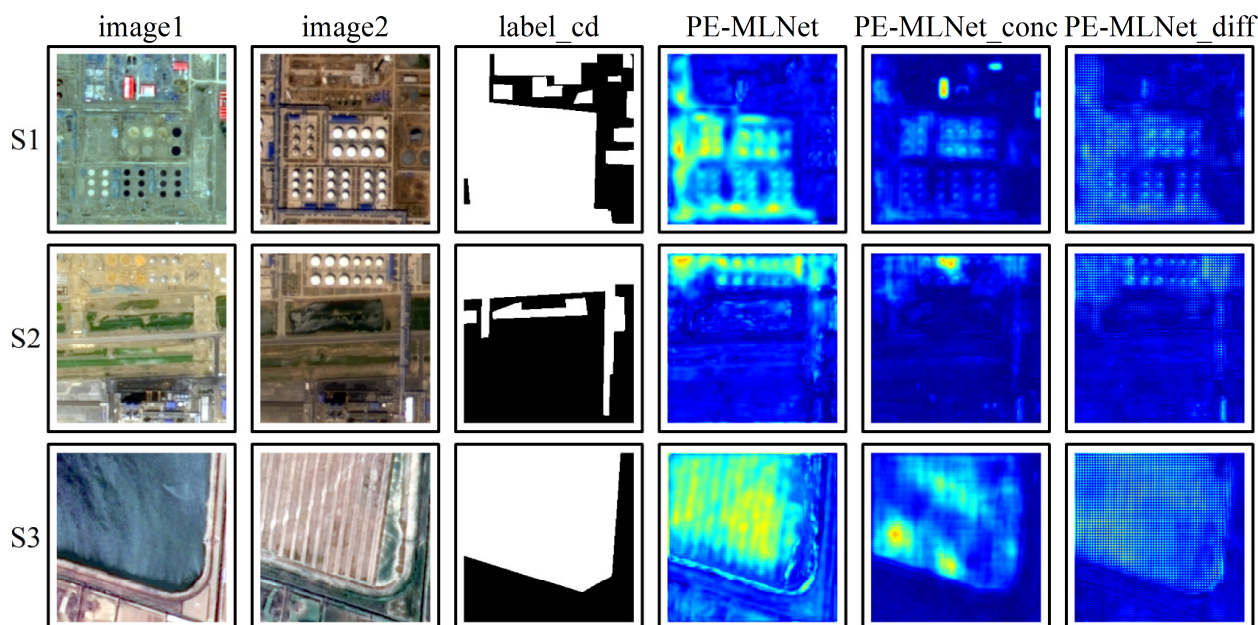


Figure 9. Visualization comparison of classification maps from different methods. The brightness indicates the magnitude of the weights; the brighter it is, the higher the weights are. S1–S3 represent three representative samples.

3.6. Human Activity Detection

By cropping high-resolution GF-6 images from 2019, 2021, and 2023 into a tile set of 256×256 pixel size and inputting them into the trained PE-MLNet model, land use changes and human activity intensity were evaluated. The detection results are shown in Figure 10. By comparing the land use types before and after the change area, different human activities can be obtained. The detailed display in Figure 11 shows that PE-MLNet can accurately identify various types of human activities, such as buildings, aquaculture ponds, roads, etc. The results of the land use transition diagrams uncovered that the areas of buildings, roads, and oil depots have increased, while the farmland area largely decreased over the five years, primarily transformed into aquaculture ponds and other land uses (Figure 12). Regarding the spatial distribution, human activities were most intense in the northeastern coastal areas, primarily reflected in the expansion of construction land and the conversion of salt fields into aquaculture ponds. In the southeast estuary region, two main types of changes were identified: The first is the degradation of wetlands, which is primarily a result of recent artificial management efforts to control the invasion of *Spartina alterniflora* in the area. The second is the extension of natural water bodies, possibly influenced by increased runoff and more precipitation than normal [49].

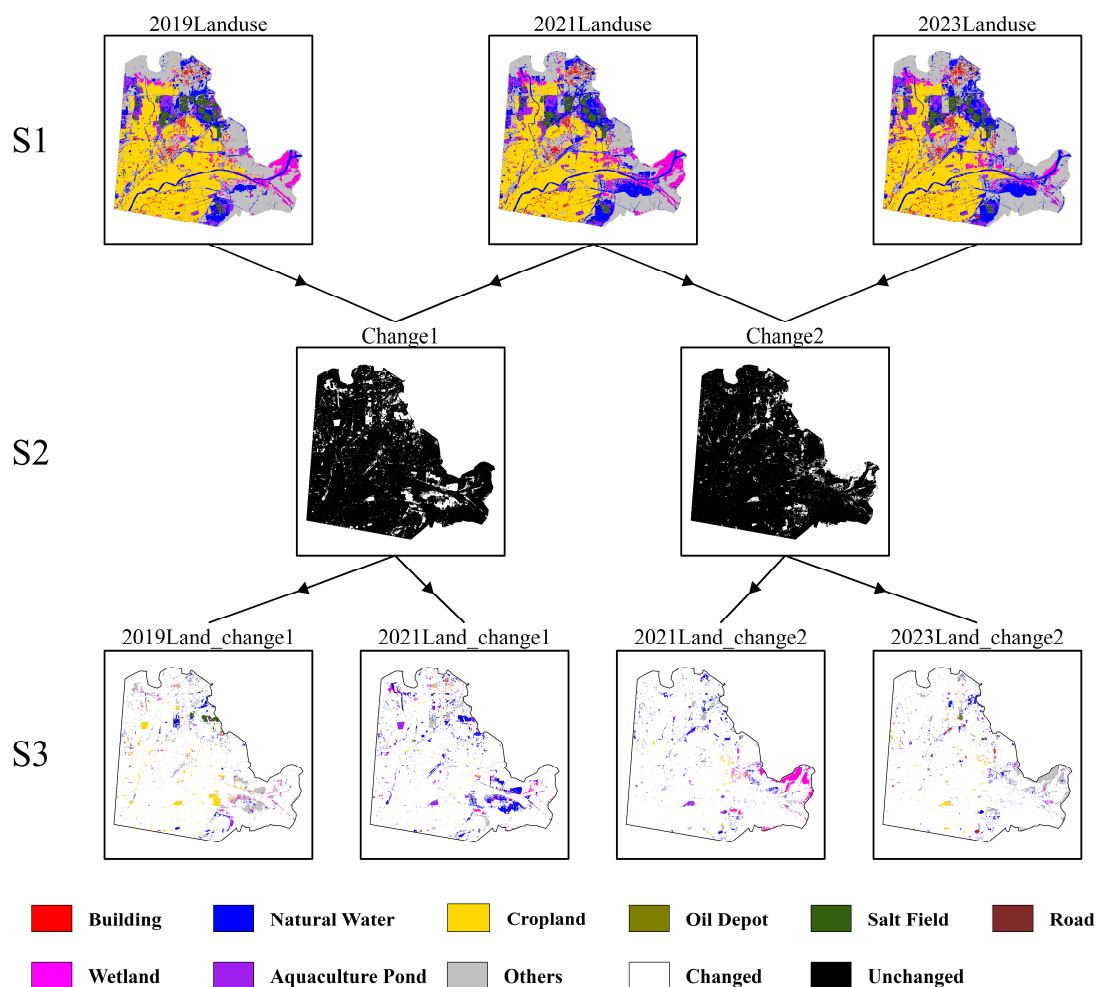


Figure 10. Human activity identification across the study area. S1 and S2, respectively, represent the land use classification map and the change detection map. S3 illustrates the land use classification before and after the change in the change area obtained by masking the classification results of S1 with S2.

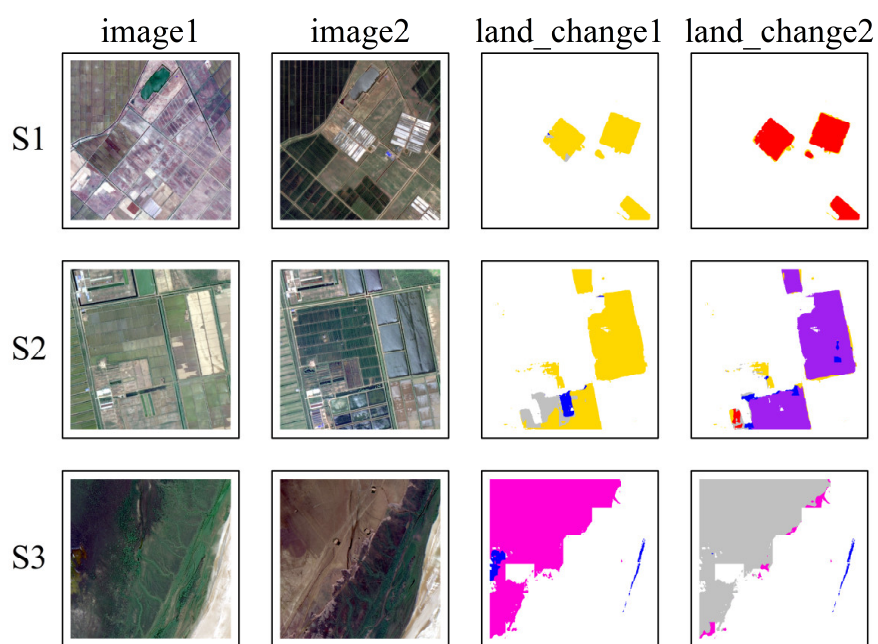


Figure 11. Cont.

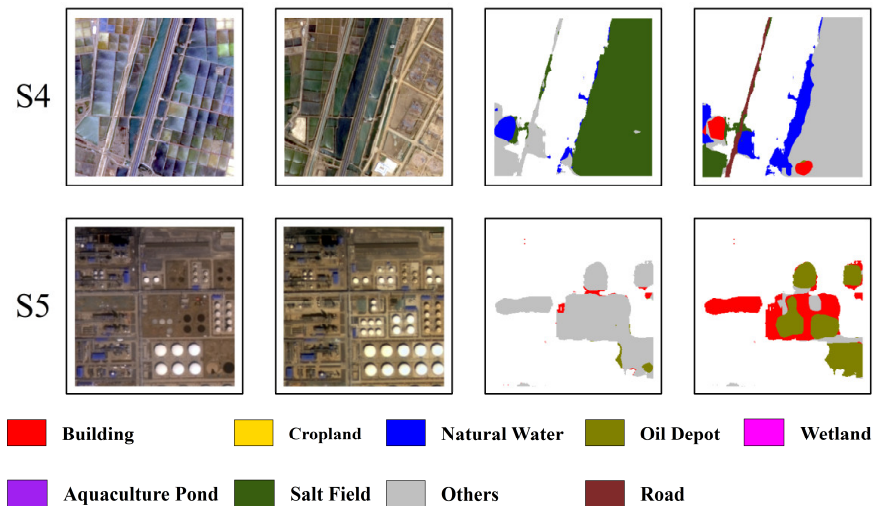


Figure 11. Detailed visualization of human activity recognition results. The *land_change1* and *land_change2* illustrate the land use classification before and after the change in the change area. S1–S5 represent five representative locations.

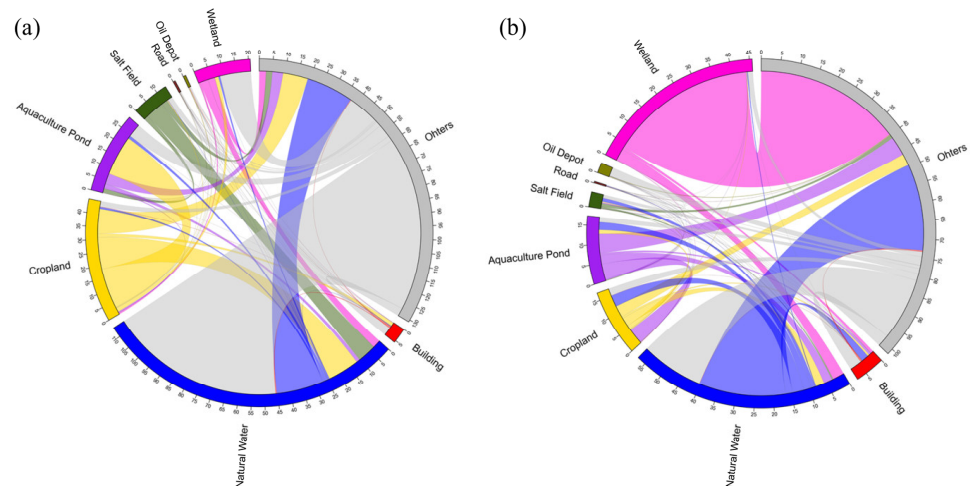


Figure 12. Land use transition diagrams (a) from 2019 to 2021 and (b) from 2021 to 2023. The color of the stripe represents the land use type being transferred out. The width of the stripe represents the size of the area, and the endpoint indicates the land use type into which it is being converted.

4. Conclusions

Rapid and intelligent detection of human activities is crucial for regional urban planning and ecological protection. High-resolution imagery provides a more detailed data source for detecting human activities. To break the limitation of most current methods only suitable for detecting specific types of human activities, this study constructed a multi-task learning-based human activity detection model named PE-MLNet to simultaneously conduct change detection and semantic segmentation in bitemporal images. By introducing the PEM to capture multi-scale change details and contextual information from bitemporal images, PE-MLNet largely outperformed other classical change detection and semantic segmentation approaches in identification accuracy and visualization details, with an average increase of 7% in the F1 score for change detection and an average improvement of 4% in the mIOU for semantic segmentation. Utilization data from the Yellow River Delta uncovered that the buildings, roads, and oil depots have obviously expanded, while the farmland area largely decreased over the five years, primarily transformed into aquaculture ponds and other land uses. In future work, we will strive to further deepen and refine the

classification system for semantic segmentation, aiming for a more granular categorization to adapt to more diverse application scenarios and needs.

Author Contributions: Conceptualization, H.L. and S.R.; methodology, H.L.; investigation, L.F.; resources, Q.W.; data curation, G.W.; writing—original draft preparation, H.L.; writing—review and editing, S.R.; supervision, Q.Z. and X.W.; project administration, J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China grant number No. 2022YFC3204400.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors are grateful to the editors and anonymous reviewers for their constructive comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wei, Y.D.; Ye, X. Urbanization, urban land expansion and environmental change in China. *Stoch. Environ. Res. Risk Assess.* **2014**, *28*, 757–765. [\[CrossRef\]](#)
2. Jiang, D.; Jones, I.; Liu, X.; Simis, S.G.H.; Cretaux, J.-F.; Albergel, C.; Tyler, A.; Spyarakos, E. Impacts of droughts and human activities on water quantity and quality: Remote sensing observations of Lake Qadisiyah, Iraq. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *132*, 104021. [\[CrossRef\]](#)
3. Pricope, N.G.; Mapes, K.L.; Woodward, K.D. Remote Sensing of Human–Environment Interactions in Global Change Research: A Review of Advances, Challenges and Future Directions. *Remote Sens.* **2019**, *11*, 2783. [\[CrossRef\]](#)
4. Bennett, M.M.; Smith, L.C. Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sens. Environ.* **2017**, *192*, 176–197. [\[CrossRef\]](#)
5. Colomina, I.; Molina, P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 79–97. [\[CrossRef\]](#)
6. Zhou, D.; Xiao, J.; Bonafoni, S.; Berger, C.; Deilami, K.; Zhou, Y.; Frohling, S.; Yao, R.; Qiao, Z.; Sobrino, J.A. Satellite Remote Sensing of Surface Urban Heat Islands: Progress, Challenges, and Perspectives. *Remote Sens.* **2018**, *11*, 48. [\[CrossRef\]](#)
7. Li, Z.L.; Wu, H.; Duan, S.B.; Zhao, W.; Ren, H.; Liu, X.; Leng, P.; Tang, R.; Ye, X.; Zhu, J.; et al. Satellite Remote Sensing of Global Land Surface Temperature: Definition, Methods, Products, and Applications. *Rev. Geophys.* **2023**, *61*, e2022RG000777. [\[CrossRef\]](#)
8. Li, S.; Cao, X. Monitoring the modes and phases of global human activity development over 30 years: Evidence from county-level nighttime light. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *126*, 103627. [\[CrossRef\]](#)
9. Brekke, C.; Solberg, A.H.S. Oil spill detection by satellite remote sensing. *Remote Sens. Environ.* **2005**, *95*, 1–13. [\[CrossRef\]](#)
10. Martin, R.V. Satellite remote sensing of surface air quality. *Atmos. Environ.* **2008**, *42*, 7823–7843. [\[CrossRef\]](#)
11. Tronin, A.A. Satellite Remote Sensing in Seismology. A Review. *Remote Sensing* **2009**, *2*, 124–150. [\[CrossRef\]](#)
12. Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 104110. [\[CrossRef\]](#)
13. Murray, N.J.; Worthington, T.A.; Bunting, P.; Duce, S.; Hagger, V.; Lovelock, C.E.; Lucas, R.; Saunders, M.I.; Sheaves, M.; Spalding, M.; et al. High-resolution mapping of losses and gains of Earth’s tidal wetlands. *Science* **2022**, *376*, 744–749. [\[CrossRef\]](#)
14. Singh, A. Review Article Digital change detection techniques using remotely-sensed data. *Int. J. Remote Sens.* **2010**, *10*, 989–1003. [\[CrossRef\]](#)
15. Wang, Z.; Wang, Y.; Wang, B.; Hu, X.; Song, C.; Xiang, M. Human Activity Detection Based on Multipass Airborne InSAR Coherence Matrix. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4013905. [\[CrossRef\]](#)
16. Weiss, M.; Jacob, F.; Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **2020**, *236*, 111402. [\[CrossRef\]](#)
17. Asokan, A.; Anitha, J. Change detection techniques for remote sensing applications: A survey. *Earth Sci. Inform.* **2019**, *12*, 143–160. [\[CrossRef\]](#)
18. Ke, L.; Lin, Y.; Zeng, Z.; Zhang, L.; Meng, L. Adaptive Change Detection with Significance Test. *IEEE Access* **2018**, *6*, 27442–27450. [\[CrossRef\]](#)
19. Massarelli, C. Fast detection of significantly transformed areas due to illegal waste burial with a procedure applicable to Landsat images. *Int. J. Remote Sens.* **2017**, *39*, 754–769. [\[CrossRef\]](#)

20. Sadeghi, V.; Farnood Ahmadi, F.; Ebadi, H. Design and implementation of an expert system for updating thematic maps using satellite imagery (case study: Changes of Lake Urmia). *Arab. J. Geosci.* **2016**, *9*, 257. [[CrossRef](#)]
21. Vázquez-Jiménez, R.; Romero-Calcerrada, R.; Novillo, C.J.; Ramos-Bernal, R.N.; Arrogante-Funes, P. Applying the chi-square transformation and automatic secant thresholding to Landsat imagery as unsupervised change detection methods. *J. Appl. Remote Sens.* **2017**, *11*, 016016. [[CrossRef](#)]
22. Tewkesbury, A.P.; Comber, A.J.; Tate, N.J.; Lamb, A.; Fisher, P.F. A critical synthesis of remotely sensed optical image change detection techniques. *Remote Sens. Environ.* **2015**, *160*, 1–14. [[CrossRef](#)]
23. Radhika, K.; Varadarajan, S.; Zhongwei, L. A neural network based classification of satellite images for change detection applications. *Cogent Eng.* **2018**, *5*, 1484587. [[CrossRef](#)]
24. Wang, Q.; Shi, W.; Atkinson, P.M.; Li, Z. Land Cover Change Detection at Subpixel Resolution With a Hopfield Neural Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 1339–1352. [[CrossRef](#)]
25. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [[CrossRef](#)]
26. Wang, J.; Bretz, M.; Dewan, M.A.A.; Delavar, M.A. Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects. *Sci. Total Environ.* **2022**, *822*, 153559. [[CrossRef](#)]
27. Bai, T.; Wang, L.; Yin, D.; Sun, K.; Chen, Y.; Li, W.; Li, D. Deep learning for change detection in remote sensing: A review. *Geo-Spat. Inf. Sci.* **2022**, *26*, 262–288. [[CrossRef](#)]
28. Bao, H.; Zerres, V.H.D.; Lehnert, L.W. Deep Siamese Network for annual change detection in Beijing using Landsat satellite data. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *130*, 103897. [[CrossRef](#)]
29. de Bem, P.; de Carvalho Junior, O.; Fontes Guimarães, R.; Trancoso Gomes, R. Change Detection of Deforestation in the Brazilian Amazon Using Landsat Data and Convolutional Neural Networks. *Remote Sens.* **2020**, *12*, 901. [[CrossRef](#)]
30. Murdaca, G.; Ricciuti, F.; Rucci, A.; Le Saux, B.; Fumagalli, A.; Prati, C. A Semi-Supervised Deep Learning Framework for Change Detection in Open-Pit Mines Using SAR Imagery. *Remote Sens.* **2023**, *15*, 5664. [[CrossRef](#)]
31. D’Addabbo, A.; Pasquariello, G.; Amodio, A. Urban Change Detection from VHR Images via Deep-Features Exploitation. In Proceedings of the Sixth International Congress on Information and Communication Technology, London, UK, 25–26 February 2021; Lecture Notes in Networks and Systems. Springer: Singapore, 2022; pp. 487–500.
32. Zhang, C.-Y.; Zhao, L.; Zhang, H.; Chen, M.-N.; Fang, R.-Y.; Yao, Y.; Zhang, Q.-P.; Wang, Q. Spatial-temporal characteristics of carbon emissions from land use change in Yellow River Delta region, China. *Ecol. Indic.* **2022**, *136*, 108623. [[CrossRef](#)]
33. Wang, Z.; Zhang, Y.; Luo, L.; Wang, N. CSA-CDGAN: Channel self-attention-based generative adversarial network for change detection of remote sensing images. *Neural Comput. Appl.* **2022**, *34*, 21999–22013. [[CrossRef](#)]
34. Zhu, Q.; Guo, X.; Deng, W.; Shi, S.; Guan, Q.; Zhong, Y.; Zhang, L.; Li, D. Land-Use/Land-Cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 63–78. [[CrossRef](#)]
35. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *arXiv* **2016**, arXiv:1612.01105. [[CrossRef](#)]
36. Huang, Y.; Li, X.; Du, Z.; Shen, H. Spatiotemporal Enhancement and Interlevel Fusion Network for Remote Sensing Images Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5609414. [[CrossRef](#)]
37. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* **2019**, arXiv:1912.01703. [[CrossRef](#)]
38. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980. [[CrossRef](#)]
39. Wu, C.; Du, B.; Cui, X.; Zhang, L. A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion. *Remote Sens. Environ.* **2017**, *199*, 241–255. [[CrossRef](#)]
40. Rosenfield, G.H.; Fitzpatrick-Lins, K. A coefficient of agreement as a measure of thematic classification accuracy. *Photogramm. Eng. Remote Sens.* **1986**, *52*, 223–227.
41. Daudt, R.C.; Saux, B.L.; Boulch, A. Fully Convolutional Siamese Networks for Change Detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
42. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [[CrossRef](#)]
43. Varghese, A.; Gubbi, J.; Ramaswamy, A.; Balamuralidhar, P. ChangeNet: A Deep Learning Architecture for Visual Change Detection. In Proceedings of the Computer Vision—ECCV 2018 Workshops, Munich, Germany, 8–14 September 2019; pp. 129–145.
44. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building Change Detection for Remote Sensing Images Using a Dual-Task Constrained Deep Siamese Convolutional Network Model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 811–815. [[CrossRef](#)]
45. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *arXiv* **2014**, arXiv:1411.4038. [[CrossRef](#)]
46. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597. [[CrossRef](#)]

47. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *arXiv* **2015**, arXiv:1511.00561. [[CrossRef](#)] [[PubMed](#)]
48. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. *arXiv* **2019**, arXiv:1902.09212. [[CrossRef](#)]
49. Qiu, H.; Yang, S.; Wang, G.; Liu, X.; Zhang, J.; Xu, Y.; Dong, S.; Liu, H.; Jiang, Z. Analysis of Carbon Flux Characteristics in Saline–Alkali Soil Under Global Warming. *J. Agron. Crop Sci.* **2024**, *210*, e12720. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.