applied sciences

MDPI

Article

Multi-Task Learning for Building Extraction and Change Detection from Remote Sensing Images

Danyang Hong [1,†], Chunping Qiu [1,†], Anzhu Yu [1,*], Yujun Quan [1], Bing Liu [2] and Xin Chen [1]

[1] The School of Surveying and Mapping, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
[2] The School of Data and Target Engineering, PLA Strategic Support Force Information Engineering University, Zhengzhou 450001, China
* Correspondence: anzhu_yu@126.com
† These authors contributed equally to this work.

Abstract: Building extraction (BE) and change detection (CD) from remote sensing (RS) imagery are significant yet highly challenging tasks with substantial application potential in urban management. Learning representative multi-scale features from RS images is a crucial step toward practical BE and CD solutions, as in other DL-based applications. To better exploit the available labeled training data for representation learning, we propose a multi-task learning (MTL) network for simultaneous BE and CD, comprising the state-of-the-art (SOTA) powerful Swin transformer as a shared backbone network and multiple heads for predicting building labels and changes. Using the popular CD dataset the Wuhan University building change detection dataset (WHU-CD), we benchmarked detailed designs of the MTL network, including backbone and pre-training choices. With a selected optimal setting, the intersection over union (IoU) score was improved from 70 to 81 for the WHU-CD. The experimental results of different settings demonstrated the effectiveness of the proposed MTL method. In particular, we achieved top scores in BE and CD from optical images in the 2021 Gaofen Challenge. Our method also shows transferable performance on an unseen CD dataset, indicating high label efficiency.

Keywords: change detection; building extraction; multi-task learning; convolutional siamese network; swin transformer; remote sensing

Citation: Hong, D.; Qiu, C.; Yu, A.; Quan, Y.; Liu, B.; Chen, X. Multi-Task Learning for Building Extraction and Change Detection from Remote Sensing Images. *Appl. Sci.* **2023**, *13*, 1037. https://doi.org/10.3390/app13021037

Academic Editors: Sicong Liu, Qiqi Zhu and Nan Wang

Received: 26 December 2022
Revised: 6 January 2023
Accepted: 9 January 2023
Published: 12 January 2023

1. Introduction

Building extraction (BE) and change detection (CD) are important yet very challenging topics in the field of earth observation. As a geospatial dense prediction task, BE is crucial for mapping, monitoring, urban management, and 3-D reconstruction [1]. Owing to the complex background and mixed pixel problems in the remote sensing (RS) imagery, BE suffers from inaccurate classification and ambiguous boundary problems. Hence, it remains difficult to satisfy practical requirements [2]. Compared to BE, building CD is an even more challenging task, as it is focused on the accurate change information prediction of buildings between images acquired at distinct intervals, rather than simply predicting buildings in one image [3]. Building CD tasks are important for applications such as urban area development and disaster management [4].

Owing to the popularity of deep learning (DL) and the advantage of its end-to-end characteristics, DL has been widely applied in computer vision [5,6], among which feature-learning-based semantic segmentation methods for both BE and CD tasks have been widely studied. Building feature representation and pixel classification are procedures commonly required for both BE and CD tasks. Improving feature representation and fusion ability is usually required to improve model performance [7–9].

For BE tasks, early convolutional neural network (CNNs)-based models, such as the fully convolutional network (FCN) [10], ResNets [11] and UNet [12], can provide

promising results with extracted rich semantic features. Incorporating more powerful backbones [13–15], sophisticated processing strategies [16,17], or semi-supervised learning [18] has improved the model performance on the BE task. However, in the method based on CNNs, the down-sampling operation will lose spatial details of the high-resolution images, resulting in blurry edges of the extracted buildings. Some methods attempt to add edge information in building extraction [19,20], which can boost the ability of a model to perceive edges. In addition, ref. [21] designs a boundary refinement network from coarse to fine, gradually improved the construction of building edges, and suppressed the irrelevant noise of underlying features under the guidance of high-level semantics.

Similar to the BE task that takes one image as input and outputs a mask layer, the building CD task is also typically regarded as a pixel-level prediction problem. The involved classification and detection sub-tasks of a CD problem are generally dispensed together, as DL allows for the end-to-end CD and avoids the negative accumulation of errors from multiple-step-based approaches such as post-classification CD. Summarily, the building CD task can be performed by fusing multiple RS images to output building changes. There are roughly three types of strategies: late, early, and hybrid fusion, depending on how the paired images are dealt with [3]. Early fusion [22] approaches concatenate multi-temporal images as one input into a network (single-feature structure), whereas late fusion approaches [23,24] separately learn mono-temporal features and later combine them as an input to the CD network (dual-feature structure) [4]. Hybrid fusion is a combination of early and late fusions, which concatenations are carried out for both the input multi-temporal images and the learned respective features.

The early CNNs employed for CD tasks were fully convolutional Siamese networks and the following variants [25]. This type of architecture features two encoding branches for feature extraction from paired input images, and one decoding branch to detect changes from feature differences. A weight-shared encoder makes it easier to detect changes. Additionally, ref. [26] introduces a global co-attention mechanism and designs an attention-guided Siamese network that is based on pyramid features and focuses on the correlation among input feature pairs. Ref. [27] combines the Siamese network and UNet network, and proposes a large-scale SCD (semantic change detection) network comprising two encoders and two decoders with shared model parameters. Similarly, ref. [28] utilizes a UNet-based Siamese network to learn representations from bi-temporal inputs via the encoder separately, and performed a difference connection to improve the generated difference maps. Furthermore, ref. [29] proposes a multi-task constrained deep Siamese convolutional network containing three sub-networks: a CD network and two dense label prediction networks, which improved the CD accuracy. In addition, generative adversarial networks (GANs) [30,31] and recurrent neural networks (RNNs) [23,24] have been studied for CD tasks. For example, ref. [24] proposes a general and new deep Siamese convolutional multi-layer RNN for CD from multi-temporal, very high-resolution (VHR) imagery via integrating the merits of CNN and RNN.

The impact of learning strategies on network performance have been demonstrated for many kinds of vision tasks and various types of learning strategies, such as the attention mechanism, multi-scale feature fusion, and transfer learning, have been proposed [32]. Based on one of the most common strategies, the attention mechanism [33,34], introduces two kinds of general attention mechanisms in the position and channel dimensions and designs a local-global pyramid network. Notably, ref. [35] proposes a novel self-attention mechanism for spatial-temporal relationship modeling, and [36] proposes a super-resolution-based CD network with stacked attention modules. In addition, ref. [37] proposes a multi-scale supervised fusion network (MSF-NET) based on attention mechanism. In addition, transfer learning employs knowledge from other data sources by fine-tuning the pre-training models from related tasks to address the problem of limited annotations. For instance, ref. [38] proposes a transfer-learning-based CD method using recurrent FCNs with multi-scale 3D filters. Additionally, ref. [39] proposes a CNN-based

CD method with a novel loss function to obtain transferable model performance among various datasets.

Compared with the above classical change detection of dual-temporal remote sensing data, ref. [40] proposes single-temporal supervised learning (STAR) for CD by utilizing land cover and land use changes in unpaired images as a kind of supervisory information. Recently, ref. [41] proposes a graph-based segmentation for multivariate time series algorithm (MTS-GS) to analyze the change in a multivariate time series by considering all variables as an entirety, rather than treating multivariate time series as univariate time series one by one like classical change detection methods. Ref. [42] is another interesting work of CD construction, which proposes a feature decomposition–optimization–recombination network based on the decoupling idea in semantic segmentation. In addition, the problem of edge refinement in building change detection is also a hot research direction in recent years. Ref. [43] proposes an end-to-end building change detection framework that integrates discriminative information and edge structure prior information into a DL framework to improve building change detection results, especially to generate better edge detection results. It is also worth noting that while BE and CD have gained considerable development in the past few decades, the updating of building databases has not been fully studied. In order to automatically update the building footprints with minimal manual labeling of the current building ground truths, ref. [44] proposes a saliency guided edge preservation network to maintain accurate building boundaries, which is used to update the existing building database to generate the latest building footprint, which is crucial for the vector-ization of building contours.

Although several change detectors have been proposed, the methods suffer from problems such as insufficient training data, which leads to model overfitting and severely limits the application of trained models. A potential solution to this problem is multi-task learning (MTL), as it can introduce more related supervising signals during network training. While BE and CD are two highly correlated topics in the RS field, most previous studies have treated them separately. On the contrary, we hope to solve these two tasks simultaneously through an MTL framework, the potential of which was demonstrated in [45]. In the MTL setting, each task influences the other, and we assume that the BE task positively promotes the building CD task.

In this study, we propose an MTL framework to simultaneously extract buildings and detect building changes from dual-time remote sensing images by taking advantage of advanced networks, including the Swin transformer [13] and Segformer [46]. The contributions of this study are as follows:

- We propose an MTL framework based on an advanced transformer-based backbone and lightweight BE and CD heads.
- We provide benchmark results to validate design details, including backbone choice and pre-training strategies of our proposed solutions using open datasets.
- We achieved a score of 81.8214 in the "BE and CD in optical images" subject of the 2021 Gaofen challenge, which is a few tenths of points behind the first place.

The reminder of this manuscript goes as follows. Section 2 presents the model choice, employed Swin-L (the large version of Swin) backbone, and BE and CD network heads. Section 3 describes the utilized datasets and experimental setup for testing the model performance in detail. Section 4 first tests the BE and CD accuracies of different models for single-task learning to select an optimized MTL setup and then displays and compares the BE and CD results from MTL based on the selected setups, and also shows the CD results from the challenge. Section 5 discusses and analyzes the benefits of MTL and its possible reasons, the validity of the pre-training weights, and the generalization of the proposed MTL approach. In the end, Section 6 provides summaries and conclusions of the study.

2. Materials and Methods

A MTL architecture, with a shared parsing network and different outputs for each task, was proposed; the modules of a Swin transformer [13] and lightweight heads were combined as illustrated in Figure 1. As the Swin transformer is the backbone to learn multi-scale features from the inputs. The following are the three branches of different tasks. For tasks based on pixel semantic information, multi-level feature maps were fused to better predict the semantic labels for every pixel. An all-multilayer perceptron (MLP) head is used to predict building labels from each of the bi-temporal inputs, and a lightweight convolution-based head is used for CD.

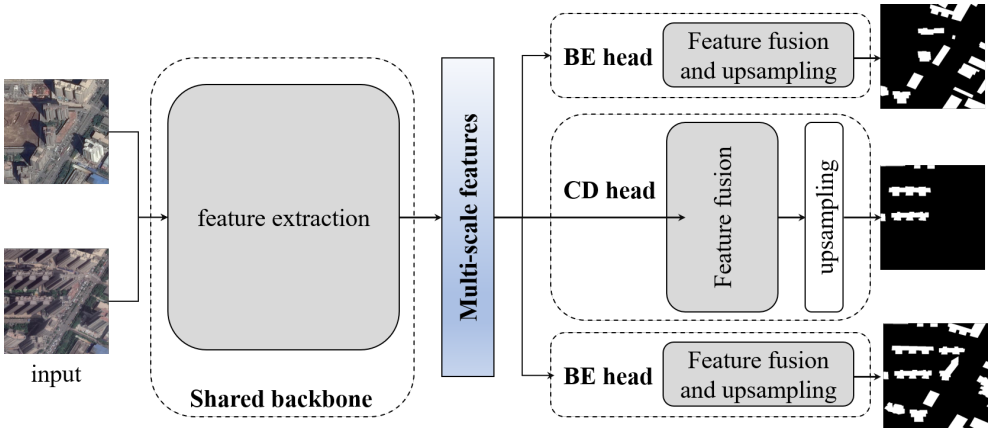

Figure 1. Overview of the proposed multi-task learnin (MTL) network with a shared backbone for multi-scale feature learning, two heads for building extraction, and one head for change detection.

The feature extraction part of our proposed MTL model is a Siamese network built on two backbones with shared weights. The backbone architecture can be selected arbitrarily, for example, ResNets. The backbone processes two input patches separately and outputs learned features in a high-dimensional feature space. The shared weights can help enhance the feature similarity between unchanged areas and reduce the feature similarity in changed areas. Subsequently, these embedded features from the backbone were used for the respective BE and CD tasks via different heads. The heads correspond to the decoder part in a common semantic segmentation network, the choices of which include UNet-like, FCN-like, and more advanced attention-based approaches.

When adding different losses of different tasks for optimization, weights are usually required, and manually tuning these weights as hyper-parameters is tedious. To balance the BE and CD tasks, we learned the weights of different tasks when combining the respective losses. We implemented the weighting of tasks on the basis of homoscedastic uncertainty, which was first introduced by [47]. The multi-task loss function used in our work can be expressed as:

$$L_{\mathrm{mt}} = \sum_{\tau}\left(\frac{1}{\sigma_{\tau}^{2}} L_{\tau}(W_{\tau}) + \log \sigma_{\tau}\right) \tag{1}$$

where $\tau \in \{BE\_before, BE\_after, CD\}$ are the three respective tasks; $L_{\tau}(W_{\tau})$ is the binary cross-entropy loss for each task; W_{τ} is the trainable parameter corresponding to each task; and σ_{τ} is a weighting parameter that affects the contribution of the individual task. The regularization term $\log \sigma_{\tau}$ avoids trivial solutions for extremely small weighting parameters. We trained the weighting terms along with the network parameters as $s := \log \sigma^{2}$ for numerical stability during the optimization. Where the network parameters is W_{τ} ,the weighting terms is s.

2.1. Backbones for Representation Learning

We employed a large version of Swin (vision transformer by shifted windows, Swin-L) as our feature extractor; this approach has the advantage of multi-scale feature modeling

flexibility from the hierarchical structure and long-range dependency encoding from the prevalent transformer architecture. As listed in Table 1, the Swin-L backbone primarily comprises four modules: a first patch partition to reduce the input patch, a linear embedding to increase the number of feature channels, several Swin transformer blocks, and patch merging in a subsequent order. The details of Swin-L can be found in [13,15]. We used features from all four stages at different scales to address large building variations.

Table 1. The main operations of the utilized large version of Swin (Swin-L) structure as the backbone.

	Layer Type	Input Size	Output Size (Equivalent Size)
	Patch partition	$512 \times 512 \times 3$	$128 \times 128 \times 3$
	Linear embedding		$16,384 \times 192$ $(128 \times 128 \times 192)$
stage 1	Swin transformer block $\times 2$	$16,384 \times 192$	$16,384 \times 192$ $(128 \times 128 \times 192)$
stage 2	Patch merging		4096×384
	Swin transformer block $\times 2$	4096×384	4096×384 $(64 \times 64 \times 384)$
stage 3	Patch merging	4096×384	1024×768
	Swin transformer block $\times 18$	1024×768	1024×768 $(32 \times 32 \times 768)$
stage 4	Patch merging	1024×768	256×1536
	Swin transformer block $\times 2$	256×1536	256×1536 $(16 \times 16 \times 1536)$

2.2. Network Heads for Building Extraction and Change Detection

BE head. The adapted all-MLP head is illustrated in Figure 2 and consists of three linear layers and one upsampling layer. This head takes the multi-scale representations learned by the backbone, that is, the Swin transformer, as inputs and outputs a segmentation mask that has the same size with the input patch. Specifically, the first of the MLP combines multi-level features from the backbone into features of the same size by resize and concatenation operations. These combined features are subsequently fused by a second layer before the third predicts the buildings and a final upsampling operation is performed to recover the resolution. The output of each branch is used for calculating the binary cross-entropy loss together with the input ground truth of building labels.

CD head. The employed CD head consists of four combinations of Convolution-BatchNorm-ReLU, progressively decreasing the feature channels, a final convolution layer to predict the changing mask, and a final upsampling layer to recover the resolution. The input to the CD head is a concatenation of the two output features from the multi-temporal inputs respectively, and the output is a one-channel mask layer indicating the changed pixels. The prediction output is used for calculating the binary cross-entropy loss together with the input ground truth during training. For inference, a sigmoid activation operation is applied to the prediction output, and the final pixel-level change map is obtained given a threshold. In our study, the threshold is set to 0.5.

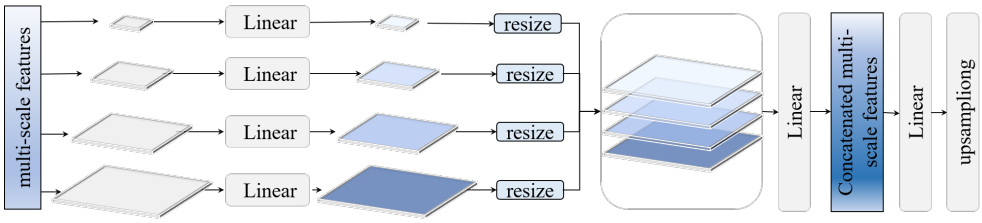

Figure 2. Architecture of the lightweight change detectio (CD) head.

3. Experimental Setup

3.1. Datasets

The Wuhan University building (WHU) CD dataset (WHU-CD) [48] is a commonly used public available dataset. The dataset includes 12796 independent buildings extracted from aerial images. The image size is 32,507 × 15,354 and the pixel spacing is 0.2 m. It covers 20.5 km² of the Christchurch area in New Zealand, in 2012 and 2016. The WHU-CD is used to benchmark different BE and CD network choices and experimental setups. The entire image is first cropped into 512 × 512 patches, and the whole dataset is then randomly split into three parts with a ratio of 7:1:2 for training, validation, and testing, respectively.

The BE and CD dataset in the Gaofen Challenge (Gaofen-BECD) included the Gaofen-2 and Jilin-1 multi-temporal satellite imagery, and the pixel spacing is smaller than 1 m. All 4000 images with a size of 512 × 512 formed 2000 paired bi-temporal images. All images provided were used as the training set. To create a validation set, approximately 10% of the training set was randomly selected and processed by random 0–10 degrees rotation, random flip, and random 0.9–1.1 scaled resize. Our training set contains 2000 paired images, and the validation set contained 200 generated paired images.

3.2. Implementation Details

We initialized Swin Backbones with the pre-trained model parameters from the ImageNet-1K dataset [13], and implement all models through the PyTorch framework on a GPU A100 developed by NVIDIA of the United States with 40 GB of memory. The ImageNet -1K dataset was founded by Feifei Li, a professor at Stanford University in the United States. During the training, we applied some often used data augmentation skills, including random rotation, random flip, and random mirror. We used the AdamW optimizer to train all networks in both the single task and multi-task settings. We set the initial learning rate to 0.001 and the learning rate scheduler uses the step learning strategy. The momentum is set to 0.9, the weight decay to 0.00001 and the batch size to 16.

For the CD task, we first initialize and froze the backbones with pre-trained weights on the BE task and only train the CD head for 5 epochs, after which all parameters are updated for 95 epochs.

3.3. Baseline Methods and Metrics

The following four state-of-the-art (SOTA) and mainstream BE and CD networks were compared with the proposed approach.

- **UNet for BE.** A standard ResNeXt101-based UNet, which is a typical network for semantic segmentation tasks, is used as a BE baseline method [12].
- **Foreground-Aware Relation Network (FarSeg) for BE.** The encoder of FarSeg consists of a ResNet-based backbone to produce pyramidal feature maps with a strategy similar to that of the feature pyramid network (FPN) and a foreground-scene relation module to improve the embedings with associating geospatial scene-relevant context. The decoder recovers the spatial size of the relation-enhanced multi-scale feature maps. The foreground-scene relation sub-network refines each level of the pyramidal feature maps using a relation map, a similarity matrix calculated with the scene embedding and the foreground representation. The geospatial scene embedding, a 1-D feature vector, was produced by an additional branch in the backbone via global context aggregation, and the foreground represents the multi-scale features.
- **Siam-UNet and Siam-UNet++ for CD.** Siamese-UNet is mainly composed by two main parts: the siamese network for feature extraction from multi-temporal inputs and the decoder for analyzing embedding differences [25]. In the feature extraction step, images of two different periods are processed by the two branches within the siamese network with shared weights. Siam-UNet ++ uses UNet++ as the backbone network, with the advantages of capturing fine-grained details by exploiting multiple nested and dense skip connections to obtain multi-scale features and reduce the pseudo-changes induced by scale variances [22].

- **ChangeStar for single task CD and MTL of BE&CD**. ChangeStar [40] consists of a dense prediction model for feature extraction and a ChangeMixin module to detect object change. The ChangeMixin module consists of a temporal swap module (TSM) and a shallow FCN involving combinations of Convolution-BatchNorm-ReLU. The TSM module takes bi-temporal feature maps from the backbone (FarSeg) as input, which are then concatenated in the channel axis in two different temporal orders; the two respective outputs from TSM are subsequently used as inputs to two FCNs with shared weights. Notably, the ChangeStar can be constructed for simultaneous BE and CD if a regression module is introduced for BE probability estimation using the extracted semantic features. Here, we use the FarSeg as the semantic segmentation model, which helps achieve the best CD result in [40].

All of these baselines employ a ResNeXt101-based backbone for a fair comparison. For the same reason, all models in this study ends with a upsampling layer with a ratio of four to recover the resolution of the prediction.

To be consistent with previous studies, we assess different approaches via four metrics: IoU, precision, recall, and *F1*-score, with respect to the building and the changed category. We do not use mean IoU for model evaluation as the number of different classes is very unbalanced, and usually the building and the changed pixels are of interest.

The challenge uses a pixel-level evaluation metric, the *F1* score, which is widely used and calculated using the following equation:

$$F_1 = \frac{2 \times P \times R}{P + R}.$$

P (precision) and R (recall) are calculated using the following two equations:

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN},$$

in which TP are true positives, FP are false positives, FN are false negatives. Assuming that F_1_before and F_1_after represent the scores of the BE results of the paired images, and F_1_change represents the CD results, the final score for the Gaofen Challenge is calculated by $Score = 0.2 \times (F_1\_before + F_1\_after) + 0.6 \times F_1\_change$.

4. Experimental Results

The quantitative and qualitative BE and CD results are presented and compared in this section. To determine the optimal settings for the MTL, we first carried out a series of ablation studies for each task in the single-task setting. Based on the results, we performed MTL experiments to improve the model performance and validate our MTL method.

4.1. Ablation Studies for Optimal Multi-Task Learning Setups

WHU-CD is used to benchmark both the BE and CD tasks using the SOTA methods.

4.1.1. Building Extraction via Single Task Learning

Table 2 lists the BE results of the SOTA methods, where our adapted Swin-based approach provides the best performance for all four metrics, followed by FarSeg and UNet, with quite high accuracy.

Table 2. Building extraction (BE) results of three different networks. The bold values represent the best results.

Network	IoU	Precision	Recall	F1
UNet	89.89	96.42	93.00	94.68
FarSeg	90.52	97.02	93.12	95.03
Ours	**91.65**	**97.19**	**94.15**	**95.65**

4.1.2. Change Detection via Single Task Learning

Table 3 lists the CD results from the six SOTA methods with four different network architectures and two backbone pre-training choices. It can be observed that our adapted Swin-based model outperforms Siam-UNet, Siam-UNet++, and ChangeStar. Additionally, the pre-training backbones on the WHU-CD are much better than those on ImageNet.

4.2. Quantitative Assessment of Multi-Task Learning

Table 4 lists the MTL results for ChangeStar and the proposed methods. The CD head was kept constant when testing the ChangeStar-based approach. The backbones for both approaches were first pre-trained on the BE task using the WHU-CD, as the comparisons in Section 4.1.2 have shown the advantages of the dataset.

Our proposed MTL idea can significantly improve CD performance, with IoU improving from 72.44 to 81.86 and from 70.46 to 78.33 for our model and ChangeStar, respectively.

Table 3. CD results of different models using pre-training choices. The bold values represent the best two results.

Model	Pretrain	IoU	Precision	Recall	F1
Siam-UNet	ImageNet	59.77	78.80	71.23	74.82
Siam-UNet++	ImageNet	62.07	71.67	**82.25**	76.59
ChangeStar	ImageNet	62.20	73.59	80.08	76.70
Ours	ImageNet	64.15	**84.89**	72.10	78.16
ChangeStar	WHU-CD	**70.46**	78.84	**86.88**	**82.67**
Ours	WHU-CD	**72.44**	**86.74**	81.47	**84.02**

Table 4. Comparative MTL results from ChangeStar and ours.

Model		IoU	Precision	Recall	F1
ChangeStar	BE	90.00	97.14	92.44	94.73
	CD	78.33	89.03	86.70	87.85
Ours	BE	91.38	97.21	93.85	95.50
	CD	81.86	91.27	88.81	90.02

4.3. Qualitative Assessment

Figures 3 and 4 present some examples of the CD results from single-task learning and MTL, corresponding to Tables 3 and 4, respectively. These figures can be used to compare the false negatives and false positives of the different approaches.

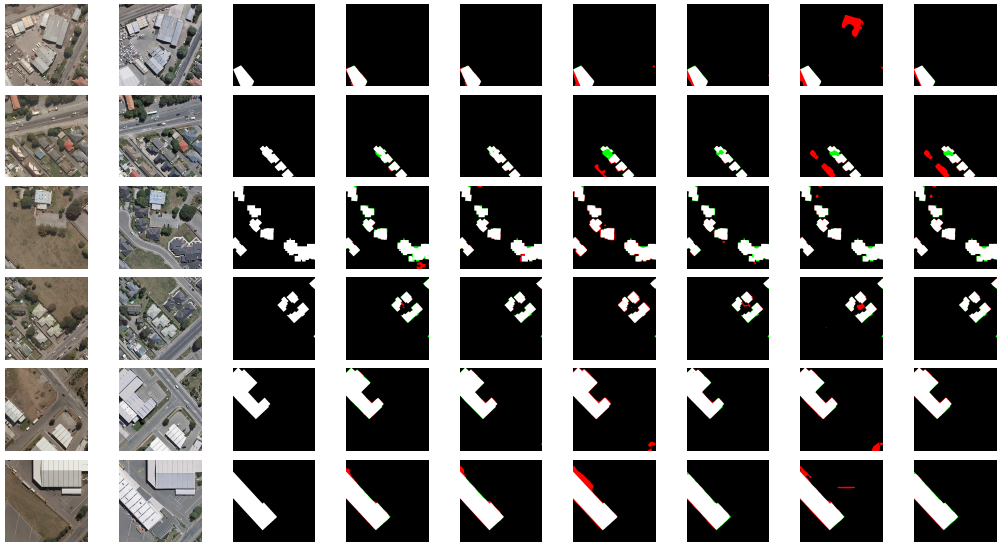

Figure 3. Comparative CD results of different models and pre-training choices on the Wuhan University building CD dataset (WHU-CD) testing set. (**a**) t_1 image patch. (**b**) t_2 image patch. (**c**) ground truth (GT) patch. (**d**) Results of Siamese-UNet pre-trained on the ImageNet dataset. (**e**) Results of Siamese-UNet++ pre-trained on the ImageNet dataset. (**f**) Results of FarSeg pre-trained on the ImageNet dataset. (**g**) Ours pre-trained on the ImageNet dataset. (**h**) Results of FarSeg pre-trained on the WHU-CD dataset. (**i**) Ours pre-trained on the WHU-CD dataset. Red indicates false positives, and green indicates false negatives.

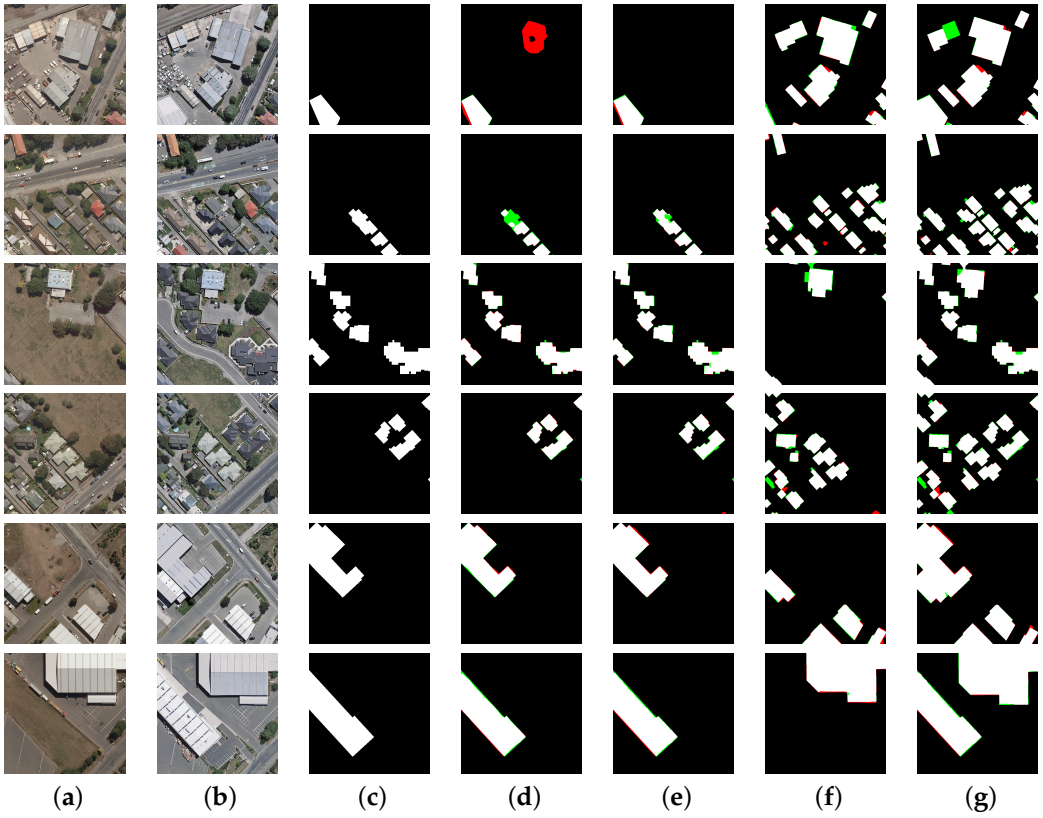

Figure 4. Comparative BE and CD results of two models on the WHU-CD testing set. (**a**) t_1 image patch. (**b**) t_2 image patch. (**c**) CD GT patch. (**d**) CD results of ChangeStar. (**e**) Our CD results. (**f**) Our BE results from t_1 Image. (**g**) Our BE results from t_2 Image. Red indicates false positives, and green indicates false negatives.

4.4. Experimental Results: Fourth Place in the 2021 Gaofen Challenge

For this challenge, we exploited strategies, including test time augmentation and model ensembling, to further boost the model performance for the proposed MTL idea. Additionally, we applied a simple and effective post-processing strategy to optimize the predicted results. A threshold was selected to filter out buildings or small changes in size. We experimentally adjusted the minimal polygon size to 15 pixels, which is suitable for both BE and CD tasks.

Figure 5 presents a comparison between our approach and ChangeStar on the Challenge dataset. In this case, both approaches can provide satisfactory results because the available samples are quite similar with a compact data distribution. However, our approach outperformed ChangeStar in achieving higher scores on the test samples when submitted to the website.

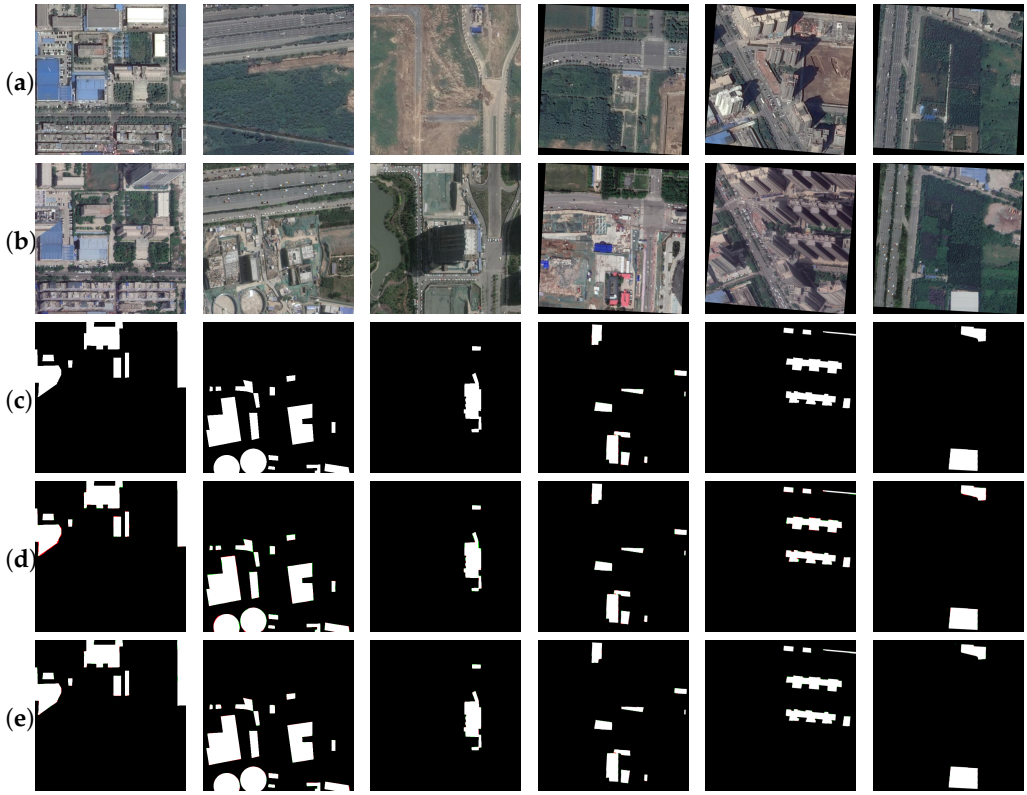

Figure 5. Comparative CD results of two models on the validation set of the BE and CD dataset in the Gaofen Challen (Gaofen-BECD). (**a**) t_1 Image. (**b**) t_2 Image. (**c**) CD GT. (**d**) CD results of ChangeStar. (**e**) Our CD results. Red indicates false positives, and green indicates false negatives.

5. Discussion

In this section, we analyze why MTL can help improve model performance and the effect of different backbones and pre-training choices. We also tested the possibility of transferring the trained model to an unseen area. Finally, we provide comments on related topics based on our observations from this study.

5.1. Benefits of Multi-Task Learning

By comparing the CD results in Table 4 with those in Table 3, it becomes clear that MTL can improve the CD results, with a significant increase in all four metrics, for both ChangeStar and our Swin-based networks. This demonstrates the benefits of the MTL for the CD task. One possible reason is that more semantic supervision signals were introduced in the MTL process, even though the available samples were the same. Increased supervision leads to higher model capacity. Additionally, the CD task in our study is an

actual building CD task, which means that the changes can only occur in building areas. Specifically, there is only one image containing buildings in the changing area. This knowledge and rich category information within the learned features are wasted when BE and CD tasks are treated separately.

We can also see that the MTL idea is so helpful that the ChangeStar-based MTL CD results (78.33 IoU) are much better than those of single task learning by Swin-L (72.44 IoU). This indicates that the advantage of a complex model architecture is less important, e.g., than domain knowledge, when it comes to a specific RS application and a fixed dataset.

In addition, MTL can result in a slight performance decrease for the BE task, as is shown by the comparative results listed in Tables 2 and 4, which is probably because the model focuses less on the easy BE task while optimizing BE and CD tasks together. This problem can be avoided by adding additional training or fine-tuning strategies.

5.2. Effectiveness of Pre-Trained Weights for CD

From Table 3 we find that the pre-training backbones are important. In addition, using the BE task and RS dataset is more helpful than using a classification task on the ImageNet dataset. As most model parameters are from the backbone, pre-training stabilizes the MTL process. This initialization makes a significant difference for both ResNet-based ChangStar and the adopted transformer-based Swin-L. It should be noted that Swin-L is initialized using pre-trained model parameters on the ImageNet dataset before training in this study, owing to training difficulty.

5.3. Transfer Performance to a Unseen CD Dataset

To further evaluate our approach in the domain shift scenario, we applied the trained models to an unseen CD dataset, LEVIR-CD [35]. LEVIR-CD is a commonly used CD benchmark dataset comprising 637 paired VHR Google Earth image patches. The patch size is 1024×1024 pixels, and the ground spacing distance is 0.5 m. Due to the construction growth, there are obvious land-cover changes in the bi-temporal imagery. Figure 6 shows a comparison of our approach's transfer performance to the strong baseline ChangeStar trained from the Challenge dataset, and Table 5 presents a quantitative comparison in such a setting, both of which demonstrate the outperformance of our proposed method.

5.4. Comments and Further Improvements

In the previous literature using WHU-CD for algorithm evaluation, different splits and data pre-processing steps, for example, different patch sizes and cropping methods, are used [49–52]. Thus, it is difficult to fairly compare the different approaches in related studies. In Section 4.1, we compare our approach to some typical models to demonstrate their performance in a single-task learning setup, after which we show the benefits of MTL to further improve the CD accuracies in Section 4.2.

We observed that satellite images from Challenge were not orthophotos, and high buildings were downwards to the side, as shown in Figure 5. This observed phenomenon also happens in other BE and CD datasets, leading to some label noises, which makes it difficult to acquire sharp boundaries for building roofs or footprints. This phenomenon makes the BE and CD approaches infeasible for some subsequent specific applications such as stereo matching and 3D reconstruction of buildings.

One limitation of our proposed MTL idea is that it requires BE labels in addition to CD labels, as both the BE and CD branches are optimized in a supervised manner, where annotations are needed during the training process. In reality, CD datasets do not necessarily have BE labels. Therefore, further improvements are also needed for the proposed method. In the future, we can either simulate pseudo CD labels using simulated paired BE samples, as carried out in [40], or resort to self-supervised learning techniques such as those proposed in [53,54] for representation learning to decrease the amount of required labels.

Table 5. Comparative CD results on the LEVIR dataset using models trained on the Challenge dataset. The bold values represent the better results.

Model	IoU	Precision	Recall	F1
ChangeStar	56.16	**69.11**	74.98	71.92
Ours	**58.65**	63.49	**88.50**	**73.94**

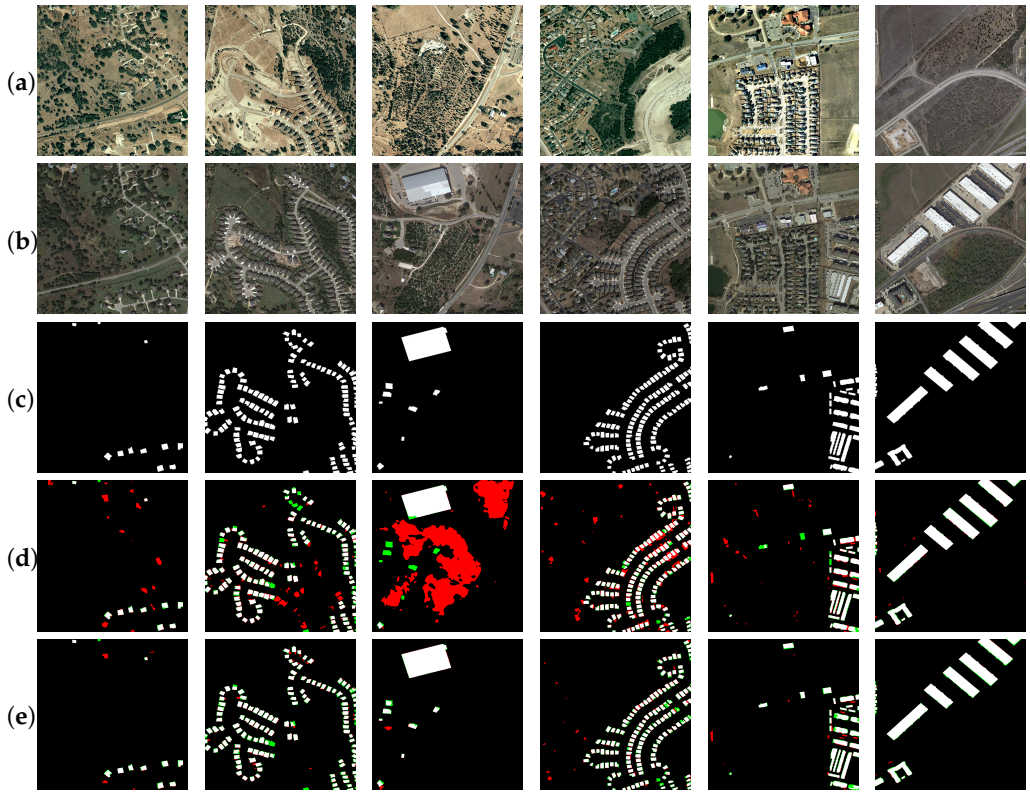

Figure 6. Comparative CD results on the LEVIR dataset of two models trained with the Challenge dataset. (**a**) t_1 Image. (**b**) t_2 Image. (**c**) CD GT. (**d**) CD results of ChangeStar. (**e**) Our CD results. Red indicates false positives, and green indicates false negatives.

6. Conclusions and Outlook

Extracting buildings and detecting their changes using VHR RS images are crucial tasks that have attracted increasing attention. The powerful SOTA DL models have achieved promising results for these challenging tasks. However, most studies treat these two tasks separately with different focuses and dedicated datasets, which leads to the redundant research effort and annotation burden. In contrast, we proposed an MTL approach for BE and CD from RS images. Specifically, we integrated a SOTA Swin transformer and light heads into the MTL network to use the available training samples for representative feature learning. In addition, we adaptively learned the task weights to balance different losses and obtain efficient and effective optimization. Extensive experimental results demonstrate the superior performance of our proposed solution than the SOTA BE and CD models using different public datasets. Moreover, we achieved fourth place in the 2021 Gaofen Challenge. We also demonstrated our approach's potential in test settings with a domain shift. Moreover, our approach can be easily adapted to a wide variety of application scenarios, such as urban monitoring from time series and general CD tasks from RS images. More related RS applications will be explored in the future to validate the proposed MTL idea.

Author Contributions: Conceptualization, A.Y. and C.Q.; methodology, A.Y.; software, A.Y. and D.H.; validation, D.H., A.Y. and C.Q.; formal analysis, D.H., A.Y. and C.Q.; investigation, A.Y.; resources, A.Y.; data curation, A.Y., X.C.; writing—original draft preparation, D.H., A.Y. and C.Q.; writing—review and editing, D.H., A.Y., C.Q., X.C. and B.L.; visualization, D.H. and Y.Q.; supervision, A.Y. and B.L.; project administration, A.Y.; funding acquisition, A.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China OF FUNDER grant No. 42101458, No. 4180138, No. 42201513.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

Abbreviations

The following abbreviations are used in this manuscript:

SOTA	state-of-the-art
BE	building extraction
CD	change detection
RS	remote sensing
MTL	multi-task learning
WHU-CD	the Wuhan University building (WHU) change detection dataset
DL	deep learning
CNN	the convolutional neural network
FCN	the fully convolutional network
SCD	semantic change detection
RNN	the recurrent neural network
GAN	the generative adversarial network
VHR	very high-resolution
MLP	multilayer perceptron
Swin_L	the large version of Swin
GT	ground truth
TSM	temporal swap module
FarSeg	Foreground-Aware Relation Network
FPN	feature pyramid network

References

1. You, Y.; Cao, J.; Zhou, W.Y. A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios. *Remote Sens.* **2020**, *12*, 2460. [CrossRef]
2. Thorsten, H.; Claudia, K. Object detection and image segmentation with deep learning on earth observation data: A review-part I: Evolution and recent trends. *Remote Sens.* **2020**, *10*, 1667.
3. Wen, D.; Huang, X.; Bovolo, F.; Li, J.; Ke, X.; Zhang, A.; Benediktsson, J.A. Change Detection From Very-High-Spatial-Resolution Optical Remote Sensing Images: Methods, applications, and future directions. *IEEE Geosci. Remote Sens. Mag.* **2021**, *4*, 68–101. [CrossRef]
4. Huiwei, J.; Min, P.; Yuanjun, Z.; Haofeng, X.; Zemin, H.; Jingming, L.; Xiaoli, M.; Xiangyun, H. A Survey on Deep Learning-Based Change Detection from High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *7*, 1552.
5. Xia, C.; Pan, Z.; Li, Y.; Chen, J.; Li, H. Vision-based melt pool monitoring for wire-arc additive manufacturing using deep learning method. *Int. J. Adv. Manuf. Technol.* **2022**, *120* , 551–562. [CrossRef]
6. Li, W.; Zhang, L.; Wu, C.; Cui, Z.; Niu, C. A new lightweight deep neural network for surface scratch detection. *Int. J. Adv. Manuf. Technol.* **2022**, *123*, 1999–2015. [CrossRef]
7. Sicong, L.; Yongjie, Z.; Qian, D.; Lorenzo, B.; Alim, S.; Xiaohua, T.; Yanmin, J.; Chao, W. A Shallow-to-Deep Feature Fusion Network for VHR Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5410213.
8. Yongjie, Z.; Sicong, L.; Qian, D.; Hui, Z.; Xiaohua, T.; Michele, D. A Novel Multitemporal Deep Fusion Network (MDFN) for Short-Term Multitemporal HR Images Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10691–10704.

9. Sicong, L.; Yongjie, Z.; Qian, D.; Alim, S.; Xiaohua, T.; Michele, D. A Novel Feature Fusion Approach for VHR Remote Sensing Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 464–473.
10. Nicolas, A.; Bertrand, L.S.; Sébastien, L. Joint Learning from Earth Observation and OpenStreetMap Data to Get Faster Better Semantic Maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1552–1560.
11. Shiqing, W.; Shunping, J. Graph Convolutional Networks for the Automated Production of Building Vector Maps From Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11.
12. Olaf, R.; Philipp, F.; Thomas, B. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241.
13. Ze, L.; Yutong, L.; Yue, C.; Han, H.; Yixuan, W.; Zheng, Z.; Stephen, L.; Baining, G. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–18 October 2021; pp. 10012–10022.
14. Xin, C.; Chunping, Q.; Wenyue, G.; Anzhu, Y.; Xiaochong, T.; Michael, S. Multiscale feature learning by transformer for building extraction from satellite images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5.
15. Chunping, Q.; He, L.; Wenyue, G.; Xin, C.; Anzhu, Y.; Xiaochong, T.; Michael, S. Transferring transformer-based models for cross-area building extraction from remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 4104–4116.
16. Weijia, L.; Conghui, H.; Jiarui, F.; Juepeng, Z.; Haohuan, F.; Le, Y. Semantic segmentation-based building footprint extraction using very high-resolution satellite images and multi-source GIS data. *Remote Sens.* **2019**, *11*, *403*.
17. Zhuo, Z.; Yanfei, Z.; Junjue, W.; Ailong, M. Foreground-Aware Relation Network for Geospatial Object Segmentation in High Spatial Resolution Remote Sensing Imagery. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 4095–4104.
18. Jian, K.; Zhirui, W.; Ruoxin, Z.; Xian, S.; Ruben, F.; Antonio, P. PiCoCo: Pixelwise Contrast and Consistency Learning for Semisupervised Building Footprint Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10548–10559.
19. Liu, Y.; Chen, D.; Ma, A.; Zhong, Y.; Fang, F.; Xu, K. Multiscale U-Shaped CNN Building Instance Extraction Framework With Edge Constraint for High-Spatial-Resolution Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 6106–6120. [CrossRef]
20. Yang, G.; Zhang, Q.; Zhang, G. EANet: Edge-Aware Network for the Extraction of Buildings from Aerial Images. *Remote Sens.* **2021**, *12*, 2161. [CrossRef]
21. Guo, H.; Du, B.; Zhang, L.; Su, X. A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery. *Isprs J. Photogramm. Remote Sens.* **2022**, *183*, 40–252. [CrossRef]
22. Daifeng, P.; Yongjun, Z.; Haiyan, G. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sens.* **2019**, *11*, 1382.
23. Lichao, M.; Lorenzo, B.; Xiao, Z.X. Learning Spectral-Spatial-Temporal Features via a Recurrent Convolutional Neural Network for Change Detection in Multispectral Imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 924–935.
24. Hongruixuan, C.; Chen, W.; Bo, D.; Liangpei, Z.; Le, W. Change Detection in Multisource VHR Images via Deep Siamese Convolutional Multiple-Layers Recurrent Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 1–17.
25. Rodrigo, D.C.; Bertrand, S.L.; Alexandre, B. 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.
26. Huiwei, J.; Xiangyun, H.; Kun, L.; Jinming, Z.; Jinqi, G.; Mi, Z. PGA-SiamNet: Pyramid Feature-Based Attention-Guided Siamese Network for Remote Sensing Orthoimagery Building Change Detection. *Remote Sens.* **2020**, *12*, 484.
27. Daifeng, P.; Lorenzo, B.; Yongjun, Z.; Haiyan, G.; Pengfei, H. Scdnet: A Novel Convolutional Network For Semantic Change Detection In High Resolution Optical Remote Sensing Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *103*, 102465.
28. Xin, Y.; Lei, H.; Yongmei, Z.; Yunqing, L. MRA-SNet: Siamese Networks of Multiscale Residual and Attention for Change Detection in High-Resolution Remote Sensing Images. *Remote Sens.* **2021**, *13*, 4528.
29. 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]
30. Lebedev, A.M.; Vizilter, Y.; Vygolov, V.O.; Knyaz, A.V.; Rubis, Y.A. Change Detection in Remote Sensing Images Using Conditional Adversarial Networks. *ISPRS- Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *42*, 565–571. [CrossRef]
31. Sudipan, S.; Francesca, B.; Lorenzo, B. Building Change Detection in VHR SAR Images via Unsupervised Deep Transcoding. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 1917–1929.
32. Jian, K.; Zhirui, W.; Ruoxin, Z.; Junshi, X.; Xian, S.; Ruben, F.; Antonio, P. DisOptNet: Distilling Semantic Knowledge from Optical Images for Weather-independent Building Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15.
33. Chenxiao, Z.; Peng, Y.; Deodato, T.; Liangcun, J.; Boyi, S.; Li, H.; Guangchao, L. 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.
34. Tongfei, L.; Maoguo, G.; Di, L.; Qingfu, Z.; Hanhong, Z.; Fenlong, J.; Mingyang, Z. Building Change Detection for VHR Remote Sensing Images via Local-Global Pyramid Network and Cross-Task Transfer Learning Strategy. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17.

35. Hao, C.; Zhenwei, S. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662.
36. Mengxi, L.; Qian, S.; Andrea, M.; Da, H.; Xiaoping, L.; Liangpei, Z. Super-Resolution-Based Change Detection Network With Stacked Attention Module for Images With Different Resolutions. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–18.
37. Jiahao, C.; Junfu, F.; Mengzhen, Z.; Yuke, Z.; Chen, S. MSF-Net: A Multiscale Supervised Fusion Network for Building Change Detection in High-Resolution Remote Sensing Images. *IEEE Access* **2022**, *10*, 30925–30938.
38. Song, A.; Choi, J. Fully Convolutional Networks with Multiscale 3D Filters and Transfer Learning for Change Detection in High Spatial Resolution Satellite Images. *Remote Sens.* **2020**, *12*, 799. [CrossRef]
39. Liu, J.; Chen, K.; Xu, G.; Sun, X.; Yan, M.; Diao, W.; Han, H. Convolutional Neural Network-Based Transfer Learning for Optical Aerial Images Change Detection. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*,127–131. [CrossRef]
40. Zheng, Z.; Ma, A.; Zhang, L.; Zhong, Y. Change is Everywhere: Single-Temporal Supervised Object Change Detection in Remote Sensing Imagery. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15173–15182.
41. Wang, N.; Li, W.; Tao, R.; Du, Q. Graph-based block-level urban change detection using Sentinel-2 time series. *Remote Sens. Environ.* **2022**, *274*, 112993. [CrossRef]
42. Yuanxin, Y.; Liang, Z.; Bai, Z.; Chao, Y.; Miaomiao, S.; Jianwei, F.; Zhitao, F. Feature Decomposition-Optimization-Reorganization Network for Building Change Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 722.
43. Bai, B.; Fu, W.; Lu, T.; Li, S. Edge-Guided Recurrent Convolutional Neural Network for Multitemporal Remote Sensing Image Building Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
44. Guo, H.; Shi, Q.; Marinoni, A.; Du, B.; Zhang, L. Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images. *Remote Sens. Environ.* **2021**, *264*, 112589. [CrossRef]
45. Ying, S.; Xinchang, Z.; Jianfeng, H.; Haiying, W.; Qinchuan, X. Fine-Grained Building Change Detection From Very High-Spatial-Resolution Remote Sensing Images Based on Deep Multitask Learning. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5.
46. Enze, X.; Wenhai, W.; Zhiding, Y.; Anima, A.; Jose, A.M.; Ping, L. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Annu. Conf. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
47. Alex, K.; Yarin, G.; Roberto, C. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491.
48. Shunping, J.; Shiqing, W.; Meng, L. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 574–586.
49. Chen, H.; Qi, Z.; Shi, Z. Remote Sensing Image Change Detection with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
50. Pang, S.; Zhang, A.; Hao, J.; Liu, F.; Chen, J. SCA-CDNet: A robust siamese correlation-and-attention-based change detection network for bitemporal VHR images. *Int. J. Remote Sens.* **2021**, *43*, 1–22. [CrossRef]
51. Junkang, X.; Hao, X.; Hui, Y.; Biao, W.; Penghai, W.; Jaewan, C.; Lixiao, C.; Yanlan, W. Multi-Feature Enhanced Building Change Detection Based on Semantic Information Guidance. *Remote Sens.* **2021**, *13*, 41–71.
52. Xueli, P.; Ruofei, Z.; Zhen, L.; Qingyang, L. Optical Remote Sensing Image Change Detection Based on Attention Mechanism and Image Difference. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7296–7307.
53. Kaiming, H.; Haoqi, F.; Yuxin, W.; Saining, X.; Ross, G. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on CVPR, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
54. Chen, X.; He, K. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF conference on CVPR, Nashville, TN, USA, 20–25 June 2021; pp. 15750–15758.