remote sensing

Article

Learning Color Distributions from Bitemporal Remote Sensing Images to Update Existing Building Footprints

Zehui Wang [1,2], Yu Meng [1,*], Jingbo Chen [1], Junxian Ma [1,2], Anzhi Yue [1] and Jiansheng Chen [1]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: mengyu@aircas.ac.cn

Abstract: For most cities, municipal governments have constructed basic building footprint datasets that need to be updated regularly for the management and monitoring of urban development and ecology. Cities are capable of changing in a short period of time, and the area of change is variable; hence, automated methods for generating up-to-date building footprints are urgently needed. However, the labels of current buildings or changed areas are usually lacking, and the conditions for acquiring images from different periods are not perfectly consistent, which can severely limit deep learning methods when attempting to learn deep information about buildings. In addition, common update methods can ignore the strictly accurate historical labels of unchanged areas. To solve the above problem, we propose a new update algorithm to update the existing building database to the current state without manual relabeling. First, the difference between the data distributions of different time-phase images is reduced using the image color translation method. Then, a semantic segmentation model predicts the segmentation results of the images from the latest period, and, finally, a post-processing update strategy is applied to strictly retain the existing labels of unchanged regions to attain the updated results. We apply the proposed algorithm on the Wuhan University change detection dataset and the Beijing Huairou district land survey dataset to evaluate the effectiveness of the method in building surface and complex labeling scenarios in urban and suburban areas. The F1 scores of the updated results obtained for both datasets reach more than 96%, which proves the applicability of our proposed algorithm and its ability to efficiently and accurately extract building footprints in real-world scenarios.

Keywords: building update; image color translation; semantic segmentation

check for updates

Citation: Wang, Z.; Meng, Y.; Chen, J.; Ma, J.; Yue, A.; Chen, J. Learning Color Distributions from Bitemporal Remote Sensing Images to Update Existing Building Footprints. *Remote Sens.* **2022**, *14*, 5851. https://doi.org/10.3390/rs14225851

Academic Editor: Marc Bosch

Received: 15 October 2022
Accepted: 16 November 2022
Published: 18 November 2022

1. Introduction

With the rapid expansion and renewal of cities around the world [1], updating existing building databases has become routine for the generation of up-to-date building footprint information [2]; a routine which can contribute to sustainable urban development [3] and ecology [4]. Traditionally, a building update is performed by manually interpreting the change area and outlining the building boundaries of the change area, which is a time-consuming and labor-intensive process, especially when analytically dealing with large areas (e.g., nationwide). Therefore, automation is essential for facilitating building change detection and building database updates.

In recent years, the main challenge of the update task has been to maintain the building detection capability of the model for unchanged areas while generating accurate segmentation results for changed areas (including building additions and demolitions). There are two main approaches to the performance of building updates: one based on building extraction [5] and the other on building change detection [6]. The former is trained using pre-temporal images and labels, and the model is fine-tuned on post-temporal images to generate the latest building segmentation results. EANet [7] and SRINet [8] respectively propose an edge perception network and a spatial residual inception network based on segmentation network for the extraction of buildings. The latter involves training the change detection model and updating the existing database by detecting the changed buildings. MDESNet [9] and MCDNet [10] respectively propose a difference-enhanced module and a feature-guided module-based multitask siamese network that learns segmentation and change detection features simultaneously for building change detection. However, the above methods have both advantages and disadvantages. The building extraction method can make full use of the a priori knowledge of existing labels, but it requires a small amount of label fine-tuning for the images from the latest period to fit the data distribution, and edge inaccuracies are present in the edges of the extraction results, which are still different from the manual labels. The building change detection method can accurately anchor the change region but requires the change detection labels of the given bitemporal images, which is different from the actual application scenario. In practical applications, most cities have already established a basic building database, so the labels of the former time phase are easier to obtain. In addition, some scholars have proposed combining the advantages of the above two methods [11] and using attention modules to learn the feature-level change information of pre- and post-temporal images for the performance of building updates based on the building extraction method. However, the accuracy decreases dramatically when the image spatial resolution decreases and the building labeling scene becomes more complex. Moreover, this approach still requires manual annotations of the post-temporal images, which is time-consuming when studying large areas. In this paper, we update the historical database using the building extraction method based on pre- and post-temporal images and pre-temporal labels.

Semantic segmentation has become one of the most commonly used tools in remote sensing image object extraction tasks over the past few years. Compared with earlier shallow feature segmentation methods (thresholding [12], edge detection segmentation [13], and region segmentation [14]) and mid-level feature segmentation methods (cluster segmentation [15], Markov random field (MRF) based model segmentation [16], and hybrid feature combination segmentation [17]), semantic segmentation is able to learn deep features with high-level semantic information through layer-by-layer neural networks. Fully convolutional network (FCN) [18] is groundbreaking in the field of image segmentation as it replaces the fully connected layer with a convolutional layer for classification tasks and recovers the original input image size via deconvolutional up-sampling to classify each pixel. UNet [19] is a U-shaped encoder-decoder architecture and the presence of a skip connection between the encoder and decoder to help the decoder better recover the details of the target object. DeepLab [20] uses atrous convolution to obtain feature maps at different scales while ensuring that the perceptual field is not reduced, thus obtaining more contextual information [21]. The pyramid scene parsing network (PSPNet) [22] not only applies atrous convolution to ResNet [23] but also adds a pyramid pooling module for better multiscale contextual aggregation and global information acquisition. In recent years, upon witnessing the great success of transformers in natural language processing (NLP), many scholars have tried to introduce transformers to vision tasks. The vision transformer (ViT) [24] is a transformer architecture that was first applied to computer vision image classification tasks. Segmentation transformer (SETR) and pyramid vision transformer (PVT) are extensions of ViT in semantic segmentation. SwinTransformer [25] composes local windows, which are shifted between layers, and utilizes UpperNet as a pyramid FCN decoder. Segformer [26] employs positional encoding-free and hierarchical transformer encoders and a lightweight all-multilayer perceptron (MLP) decoder. Transformers will continue to be widely explored in the remote sensing field [27–30].

However, the main limitations of deep learning are its strong dependence on data and its extreme sensitivity. Considering the current generation of satellites with diverse types, short revisit cycle times, and large coverage areas, it cannot be assumed that the distributions of images are always similar. In addition, depending on the times and locations of the collected data, large intraclass variations may be encountered in remote sensing images. For example, significant spectral differences in vegetation can occur due to seasonal differences. Even for images taken at different times of the day, the brightness of the same object may vary significantly. In addition, due to atmospheric effects, in some cases, even images collected by the same satellite sensor may have very different radiation intensities [31], which makes the segmentation task more difficult, and image color translation can effectively solve the problem of differences in spectral features between images from different periods.

Image color translation involves the translation of the color style of the source domain to the color style of the target domain. Early color translation methods included linear and nonlinear methods [32]. The most commonly used nonlinear method is histogram matching (HM) [33] and linear methods include the image regression (IR) method [34], Reinhard method [35], and pseudo-invariant feature (PIF) method [36]. Although the above methods are frequently used in color translation, traditional methods still have limitations when addressing complex scenes and object changes in remote sensing images.

Generative adversarial networks (GANs) [37] can align the data distributions of the source and target domains with the aim of generating pseudo-source domain images that are statistically indistinguishable from the target domain images [31]. Pix2Pix [38] uses a conditional GAN to learn input-to-output image mappings that requires paired data. Some recent works have relaxed the need to dependent on image translation learning on pairs of training data. The coupled GAN (CoGAN) [39] generates distribution estimates using samples from the boundaries for learning joint data by forcing the discriminators and generators in the source and target domains to share parameters at the low level. UNIT [40] further extends the CoGAN by assuming the existence of a shared low-dimensional latent space between the source and target domains. MUNIT [41] and DRIT [42] extend this idea to multimodal image-to-image translation by assuming two potential representations, one for "style" and another for "content". Then, cross-domain image translation is performed by combining different content and style representations. DiscoGAN [43] and CycleGAN [44] overcome the corrupted semantic structure problem by a cycle consistency loss and encourage the generated pseudo-source domain images to be effectively reconstructed when mapped back to the source domain. The attention-guided GAN (AGGAN) [45] adds an attention mechanism (AM) to each generator of CycleGAN and assigns different weights to different image positions. The attention-guided color consistency GAN (ACGAN) [46] can extract the high-level features of images, reducing the color distribution differences between multitemporal remote sensing images. Cycada [47] segments the original and generated images using a classifier trained on the original data and minimizes the cross-entropy loss between the segments. Unlike existing GANs, the generator in ColorMapGAN [31] does not have any convolution or pooling layers. It learns to translate the colors of the training data to the colors of the test data by performing only one element-by-element matrix multiplication operation and one matrix addition operation. We compare the CycleGAN, UNIT, DRIT, HM, and Reinhard image color translation methods and analyze the effect of each method on the experimental results.

In this paper, CycleGAN is proposed to be applied to the image color translation process for dual-temporal remote sensing images to reduce the differences among data distributions. In addition, based on UNet with EfficientNet [48] as the encoder, we make full use of the a priori information of the existing database to segment the buildings from the latest period and can directly predict the added and demolished change areas. In addition, the image color translation method does not require relabeling of the post-temporal images for fine-tuning to fit the data distribution, which greatly saves time and costs. Finally, a post-processing update strategy is proposed to strictly retain the historical labels for unchanged areas, calculate the ratio between the intersection area of the prediction and historical labels and then set an appropriate threshold to replace the segmentation results with historical labels, which is of great significance for high-accuracy urban mapping. The main contributions of this paper are as follows.

1. Image color translation is performed on different-phase remote sensing images using CycleGAN to smoothly translate the color distribution from the source domain to the target domain in an unsupervised manner.
2. A priori information is obtained based on a historical database using UNet(EfficientNet) to update buildings (additions and demolitions) without relabeling.
3. We propose a post-processing update strategy to replace the segmentation of unchanged regions using strictly accurate historical labels to solve the problem of inaccurate prediction edges.

The rest of this paper is organized as follows. Section 2 describes the approach of this paper in detail. Section 3 verifies the effectiveness of the CycleGAN method and UNet(EfficientNet) by comparing them with other excellent image color translation methods and semantic segmentation models, respectively, and analyzes the segmentation improvement yielded by the post-processing update strategies. Section 4 discusses the ablation experiments and threshold selection for the post-processing update strategy. Finally, Section 5 summarizes the paper.

2. Methods

This section is structured as follows (Figure 1). First, the architecture of CycleGAN and its loss are described. Second, the UNet(EfficientNet) architecture and the loss are introduced. Finally, we discuss the post-processing update strategy proposed in this paper.

2.1. Image Color Translation

CycleGAN is an unsupervised image-to-image translation method that converts information from one form to another. Its principle is based on the idea of pairwise image style translation, which translates the style of an image to another and preserves the semantic information of the original image. CycleGAN uses two discriminators and two generators to implement a mutual mapping between the source domain X and the target domain Y (Figure 2a). The generator takes a source domain image as input and generates a synthesized image with the style of the target domain. The discriminator takes an image as input and tries to identify whether it is the original image or the generated image.

CycleGAN is implemented with a forward generator G that translates image x from domain X to image $G(x)$ in domain Y and another backward generator F that translates image y in domain Y to $F(y)$ in domain X. A discriminator Dx is used to determine whether the image is from domain X or generator $F(y)$, and another discriminator Dy is used to distinguish whether the image is from domain Y or generator $G(x)$. The training procedure of CycleGAN is divided into two supervised processes, the first of which is cycle consistency supervision, in which the image is translated from domain X to domain Y and then translated from domain Y to domain X again. The synthesized image $G(x)$ needs to generate $F(G(x))$ by the backward generator F to make it as close as possible to the original input x. The cycle consistency loss constrained network is used to solve the problem that the GAN cannot output the corresponding image. Similarly, the same process is adapted to Y. The forward cycle consistency process (Figure 2b) is $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and the backward cycle consistency process (Figure 2c) is $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$. The second stage is to determine whether an image is original or generated, i.e., the discriminator Dy discriminates whether the generated image $\hat{y}$ of generator G comes from domain Y. Similarly, this process is also used by the discriminator Dx in domain X. In this case, the generator uses a residual neural network (ResNet) architecture, and the discriminator is PatchGAN [38].

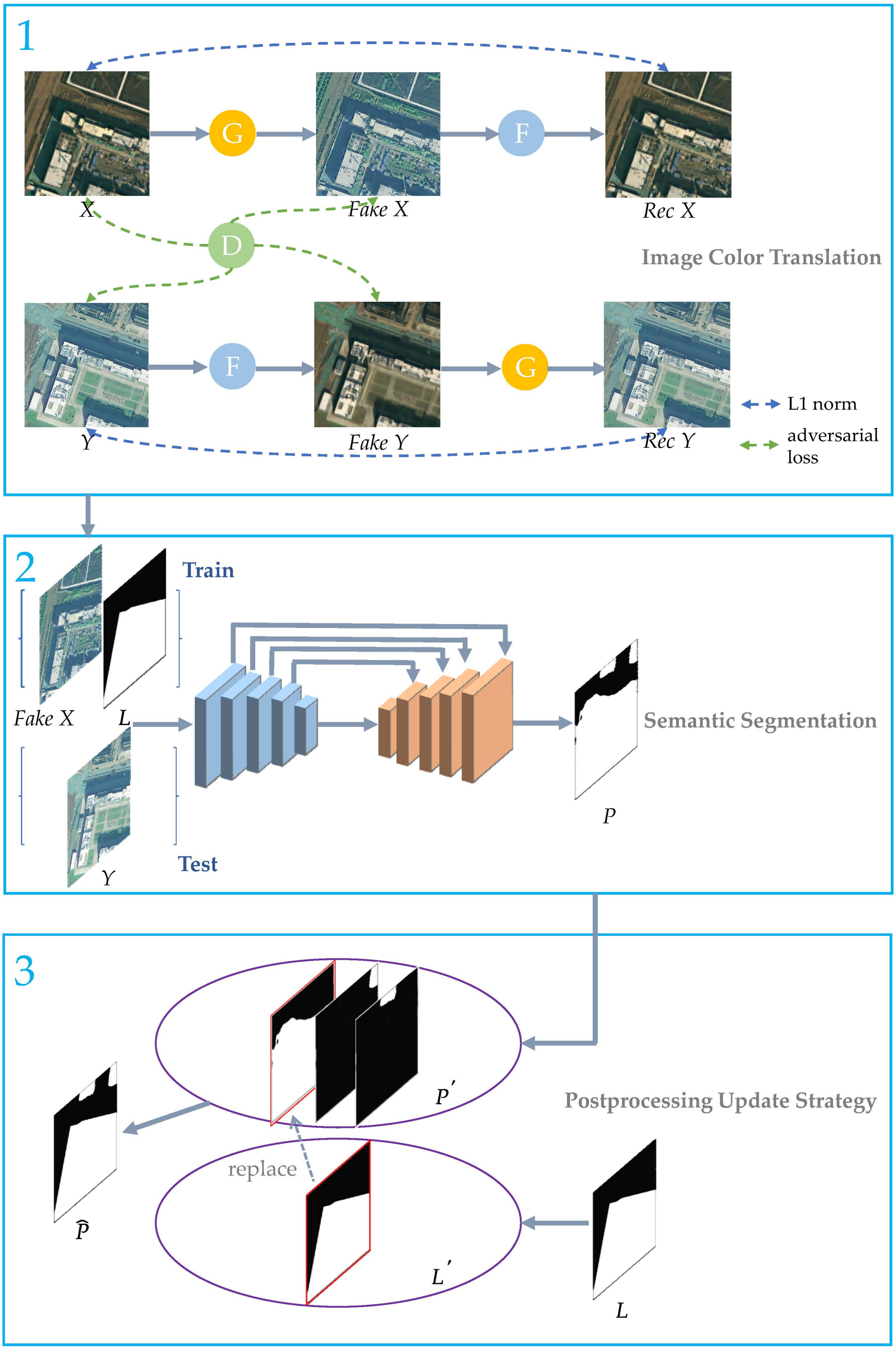


Figure 1. (**Top**) shows the image color translation method, which is used to generate pseudo-source domain images. (**Middle**) applies synthesized images to the semantic segmentation model to learn building information with the color distribution of the latest period. (**Bottom**) is the post-processing update strategy that is used to replace some predictions that meet the threshold requirement.

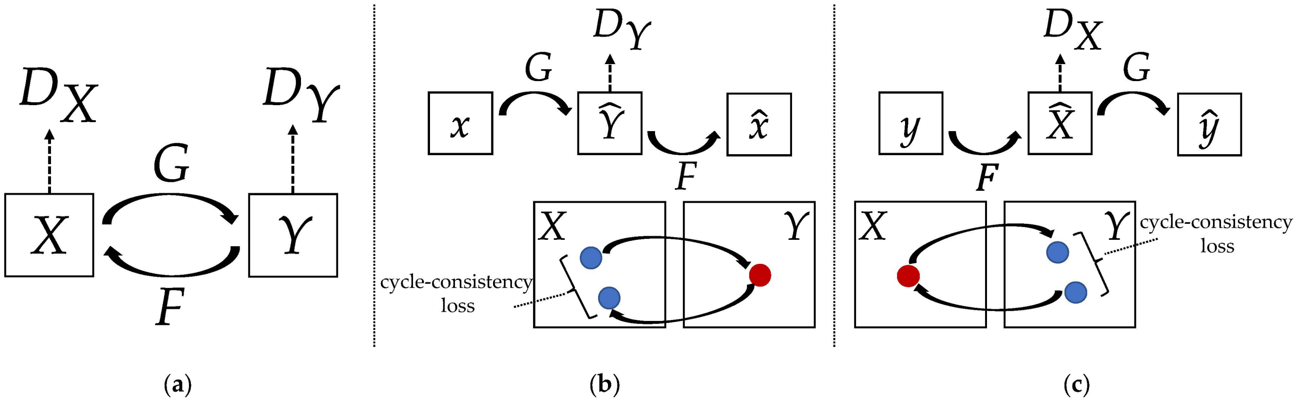


Figure 2. (**a**) The CycleGAN process for mapping and discriminating between two domains. (**b**) The forward cycle consistency process. (**c**) The backward cycle consistency process.

The losses of CycleGAN include the adversarial loss and cycle consistency loss. The adversarial loss is used to distinguish whether the generated image is real or fake. The cycle consistency loss is used to improve the ability of the model to recover the images. The forward process of the GAN involves the generator G translating the image in domain X to $G(X)$ in domain Y, and the discriminator Dy distinguishing whether it is a synthesized image. Thus, the generator G and discriminator Dy constitute the forward loss, as shown in Equation (1).

$$L_{GAN}(G, D_Y, X, Y) = E_{y\sim p_{data}(y)}[\log D_Y(y)] + E_{x\sim p_{data}(x)}\left[\log\left(1 - D_Y(G(x))\right)\right] \quad (1)$$

where $L_{GAN}(G, D_Y, X, Y)$ is the forward adversarial loss of the mapping of domain X to domain Y. The data distribution is defined as $x \sim p_{data}(x)$, $y \sim p_{data}(y)$, X denotes the source domain, Y denotes the target domain, and $E(*)$ denotes the expectation of the distribution function.

Similar to the above loss is the backward process of the GAN, where the generator F and the discriminator Dx constitute the backward loss, as shown in Equation (2).

$$L_{GAN}(F, D_X, Y, X) = E_{x\sim p_{data}(x)}[\log D_X(x)] + E_{y\sim p_{data}(y)}\left[\log\left(1 - D_X(F(y))\right)\right] \quad (2)$$

where $L_{GAN}(F, D_X, Y, X)$ is the backward adversarial loss when mapping domain Y to domain X. Both the generated image and the original image are used as inputs for discrimination.

The random combination of source and target domain images for image translation causes the model to learn different mapping relationships; therefore, relying on the adversarial loss alone does not guarantee that the function will map a single input x_i to the desired output $G(x_i)$, and there is no guarantee that the translation process will not distort the image content even for pairs of images. Therefore, to further reduce the possible mapping space and guarantee the quality of the generated images, we believe that the training process of the GAN should exhibit cycle consistency. The forward cycle consistency of CycleGAN recovers the image x of domain X to the original image x after cycle translation, i.e., $x \to G(x) \to F(G(x)) \approx x$. Similarly, for the image y of domain Y, generators G and F still satisfy backward cycle consistency: $y \to F(y) \to G(F(y)) \approx y$. The cycle consistency loss is shown in Equation (3).

$$L_{cyc}(G, F) = E_{x\sim p_{data}(x)}\left[\left|\left|F(G(x)) - x\right|\right|_1\right] + E_{y\sim p_{data}(y)}\left[\left|\left|G(F(y)) - y\right|\right|_1\right] \quad (3)$$

where the $L_{cyc}(G, F)$ formula uses the L1 paradigm.

The final loss is shown in Equation (4).

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda L_{cyc}(G, F) \tag{4}$$

where λ denotes the coefficient of the cycle consistency loss. The higher the weight is, the more important the cycle consistency loss.

2.2. Semantic Segmentation

UNet was proposed in the field of medical images and is a typical encoder–decoder architecture. The encoder uses convolution and pooling layers to increase the number of channels and reduce the spatial size to extract deep features and underlying representations of the image, and the decoder is used to recover the original size and detail information of the image. In addition, it introduces skip connections in the network to combine shallow, low-level, and fine-grained feature maps from the encoder subnetwork and deep, semantic, and coarse-grained feature maps from the decoder subnetwork [49]. Among these, the encoder uses EfficientNet [50,51], which balances the three dimensions (the network depth, width, and resolution) to capture richer, more complex, and more detailed features in images, as shown in Figure 3. Since the image color translation method can produce both the generated image *Fake X* of the pre-temporal image and the generated image *Fake Y* of post-temporal images, the generated image *Fake X* of the pre-temporal image and the pre-temporal label *L* can be used as the training data pair (*Fake X*, *L*) to learn the semantic information of the building and test for post-temporal image *Y* to obtain segmentation result *P*. Alternatively, the pre-temporal image and label can be used as the training data pair (*X*, *L*) to learn the building semantic information and to test for the post-temporal generated image *Fake Y* to obtain the segmentation result *P*. The choice of two strategies depends on the image quality between pre-temporal images and post-temporal images, usually, the original image is the upper bound of the image quality, and the better the image color translation method, the closer the generated image to this upper bound. Therefore, selecting a high-quality original image to generate the synthesized image for semantic segmentation can acquire better results. In this paper, we adopt the scheme of data pairs (*Fake X*, *L*) for training and data *Y* for testing.

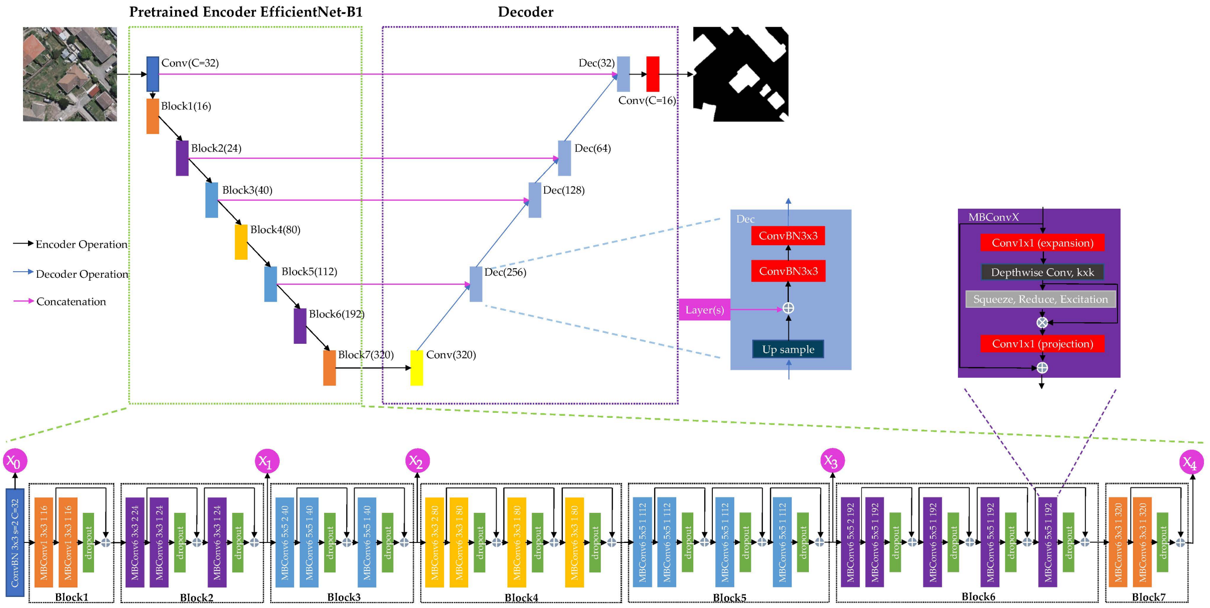


Figure 3. The top shows the UNet architecture, which consists of an encoder (green dotted box), a decoder (purple dotted box) and skip connections (pink). **The bottom** panel shows the details of the EfficientNet-b1 module, which includes seven blocks. **The right** side presents the submodule of the decoder and the MBConv module of the encoder block.

The losses for semantic segmentation include the binary cross-entropy loss L_{bce} and the Dice loss L_{dice} [52]. The binary cross-entropy loss treats each pixel as an independent sample, while the Dice loss treats it in a holistic form. The binary cross-entropy loss and Dice loss are shown in Equation (5) and Equation (6), respectively.

$$L_{bce} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\Big(p(y_i)\Big) + (1-y_i) \cdot \log\Big(1-p(y_i)\Big) \tag{5}$$

where N is the sum of all pixels in the image, y is the label, and $p(y_i)$ is the prediction probability.

$$L_{dice} = 1 - \frac{2|P \cap G|}{|P| + |G|} \tag{6}$$

where $|P \cap G|$ denotes the common number of predictions and ground truths.

The final segmentation loss is shown in Equation (7).

$$L_{seg} = L_{bce} + \gamma L_{dice} \tag{7}$$

where γ controls the importance of L_{dice} in L_{seg}.

2.3. Post-Processing Update Strategy

Since semantic segmentation is a pixel-level classification task, the central pixel is influenced by its neighboring pixels, and if the neighboring pixels belong to the interior of the target to be segmented, the central pixel classification is favored. Conversely, when the neighboring pixels are at the boundary of the target to be segmented, this negatively affects the process of correctly classifying the central pixel and thus often results in the problem of inaccurate edges. In addition, in the building update task, the changed area generally accounts for a smaller percentage than the unchanged area, so it is important to strictly maintain the corrected historical labels for urban mapping. We extract the building area contour in the prediction P and transform it into a polygon set $P' = \{p'_1, p'_2, \ldots, p'_n\}$; similarly, the pre-temporal phase label L is transformed into the set $L' = \{l'_1, l'_2, \ldots, l'_k\}$. Based on the set P', we loop through the set L'. If p' intersects with l', we can obtain the intersection region i' and then calculate the ratio $\alpha = \frac{S_{i'}}{S_{l'}}$ of the area $S_{i'}$ of the intersecting region i' to the area $S_{l'}$ of the historical label region. We set the threshold θ*; if* $\alpha > \theta$, we use l' to replace p'; otherwise, we keep p'. A larger θ indicates that the update result is more dependent on the prediction and is adapted to scenes where a longer time interval causes more change areas and a higher image resolution, and there are local changes in buildings. Conversely, a smaller θ means that the update result is more dependent on the historical label and is adapted to scenes where a shorter time interval causes fewer change areas, a lower image resolution and more complex labels. The algorithm of the post-processing update strategy is shown in Algorithm 1.

The proposed method is summarized as follows. We are given a pre-temporal image X and its corresponding label L, as well as the post-temporal image Y. The pre-temporal image X is translated into the generated image *Fake* X by the image color translation method, a semantic segmentation model is trained based on the data pair (*Fake* X, L), and the post-temporal image Y is tested to generate the building prediction P. The area of the intersection between the building prediction P and the pre-temporal label L, $S_{i'}$, is compared to the area $S_{l'}$ of the previous temporal phase label L. The building segmentation p' with a higher ratio is replaced by the corresponding building label l' from the previous temporal phase to obtain the final updated result $\hat{P}$.

Algorithm 1 Post-processing update strategy

```
Step 1:   Transform the pre-temporal label $L$ and the post-temporal prediction $P$ into the
          polygon sets $L' = \{l'_1, l'_2, \ldots, l'_k\}$ and$P' = \{p'_1, p'_2, \ldots, p'_n\}$, respectively, and set
          the threshold $\theta$.
Step 2:   Calculate $\alpha$ and update:
          for $p'$ in $P'$:
          for $l'$ in $L'$:
          if $intersection(p', l')$:
              $i' = intersection(p', l')$
              $\alpha = \frac{S_{i'}}{S_{l'}}$
          if $\alpha > \theta$: $\hat{p} = l'$
          otherwise: $\hat{p} = p'$
Step 3:   Convert the set of polygons $\hat{P} = \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n\}$ into pixel-level update results.
```

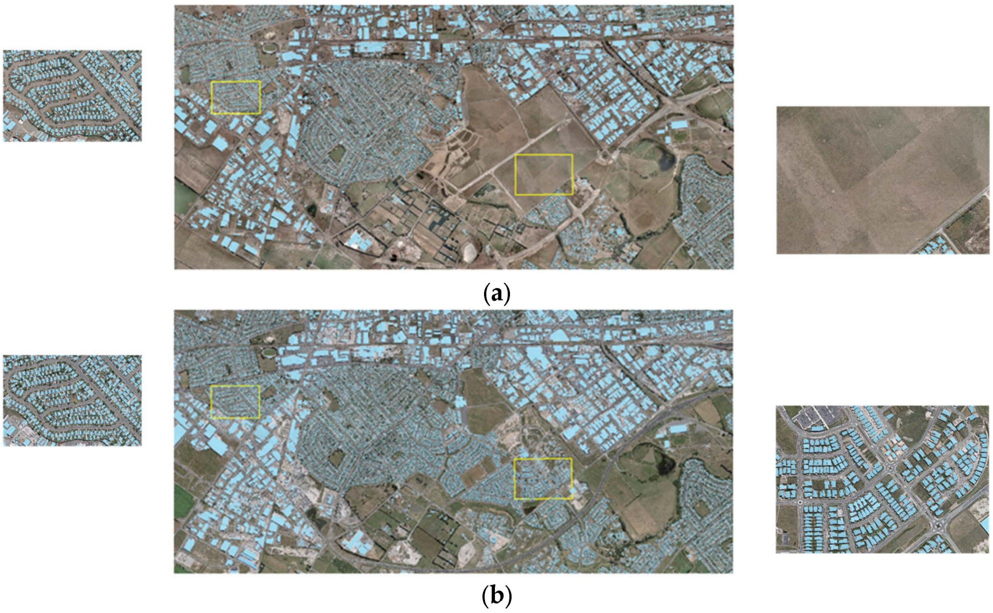
3. Experiments and Results Analysis

In this section, we first describe the two utilized datasets, introduce the implementation details of the proposed algorithm in this paper, and compare it with other excellent methods to explore the effectiveness of image color translation for the semantic segmentation of buildings from remote sensing images with different time phases. Finally, the performance of the proposed algorithm is evaluated by optimizing the segmentation process to achieve building updates using post-processing update strategies.

3.1. Datasets and Experimental Details

(1) Wuhan University Building Change Detection Dataset [53]

The study area is located in Christchurch, New Zealand, covering 20 km^2. The datasets of pre- and post-temporal images were obtained in 2012 and 2016, respectively, with a spatial resolution of 0.3 m and 3-band aerial images (Figure 4). The pre-temporal building database and post-temporal building database contain 9938 and 12,091 labels, respectively. In the pre-processing stage, the original images and building labels are cropped into patches of 256 × 256 pixels with 50% overlap, and the final total number of crops obtained is 30,107 (30,107 pre-temporal tiles for training and 30,107 post-temporal tiles for testing).


(**a**)

(**b**)

Figure 4. Wuhan University building change detection dataset, (**a**,**b**) show the images and the corresponding building labels for 2012 and 2016, respectively.

(2) Beijing Huairou district Land Survey Dataset

The study area is located in Huairou district, Beijing. The dataset contains remote sensing images and building labels from February 2018 to October 2019, covering obvious construction sites, rural settlements, soccer fields and other infrastructures with complex labeling scenarios (Figure 5). The image resolution is 2 m, and the images are 3-band images. The pre- and post-temporal databases both contain 3308 labelsand the numbers of added and demolished buildings are small due to the short interval between the two temporal settings of the land survey dataset and the large coverage of some labels. Similar to the above pre-processing output, the total number of images obtained from the final cropping operation is 8775 (8775 pre-temporal tiles for training and 8775 post-temporal tiles for testing).

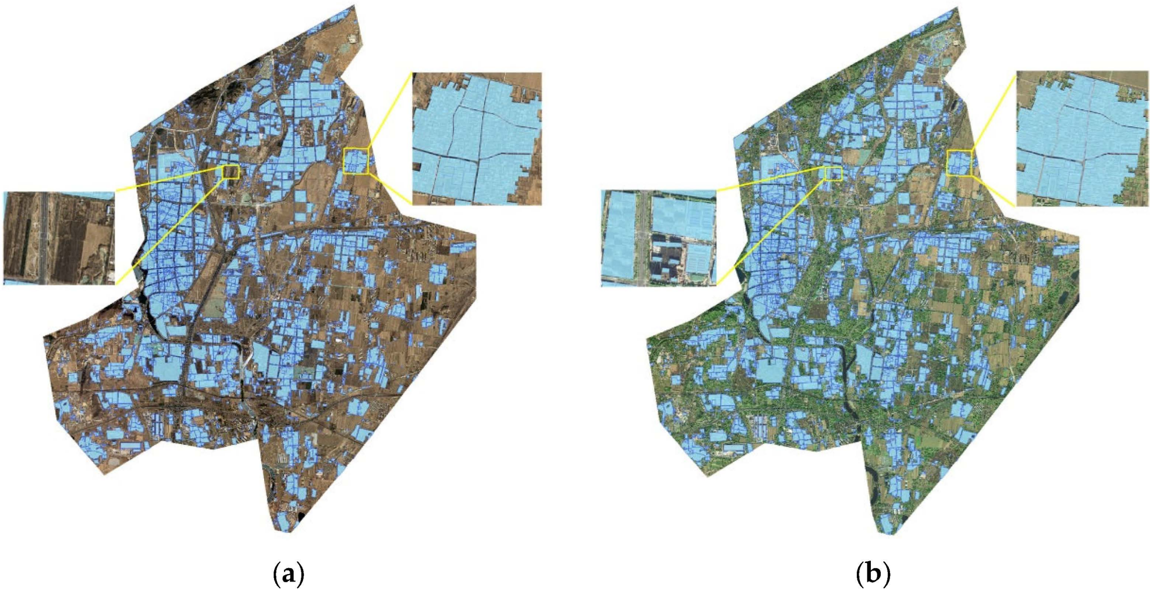


Figure 5. Beijing Huairou district land survey dataset, (**a**,**b**) show the images and the corresponding building labels for 2018 and 2019, respectively.

In the image color translation task, the generator contains three convolutional layers, nine residual blocks, two fractionally stride convolutional layers with $\frac{1}{2}$ strides and one convolutional layer that maps the feature map to RGB. The convolutional layers were followed by instance normalization [54]. The discriminator used 70 × 70 PatchGANs to discriminate whether a patch of overlapping images of size 70 × 70 was real or fake. The initial learning rate was 0.0002, the optimizer was Adam with a batch size of 4, the total number of training epochs was 100, the learning rate was linearly decayed to 0 starting from the 50th epoch, and the data were enhanced using the flipping strategy. In addition, we utilized the least-squares loss instead of the negative log loss to make the model training process more stable, generate high-quality images, and to reduce model oscillations [55]. The discriminator was updated using historically generated images instead of images generated by the current generator [56]. In the semantic segmentation task, the encoder used EfficientNet-b1 to extract image features, and the PSPNet encoder used part of the architecture of ResNet50 with a block depth of 3. The initial learning rate was 0.001, the optimizer was AdamW with a batch size of 128, the total number of training epochs was 60, and the learning rate increased linearly in the first three epochs, after which the PolyLR strategy was used to decay the rate. The data were enhanced using random cropping to 128 pixels and the flipping strategy. All experiments in this paper were performed on one NVIDIA RTX 3090 GPU.

To quantify the experimental results, five evaluation metrics, including accuracy, intersection over union (IoU), precision, recall, and F1, were used to evaluate the performance of the proposed building update method for all buildings in the whole region. First, the numbers of false-negative (FN), true-negative (TN), true-positive (TP) and false-positive (FP) pixels were calculated using the prediction and ground truth. TP indicates pixels that are correctly predicted to be positive. Conversely, FN implies pixels that are incorrectly predicted to be negative. The above evaluation metrics were then calculated auxiliary to the formulas shown below. In addition, to eliminate the differences in evaluation due to image cropping size and overlap in preprocessing, all results were calculated on the merged large map.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

$$\text{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{12}$$

3.2. Visualization of Image Color Translation

We first performed image color translation on the pre- and post-temporal images of the two datasets and evaluated the performance of each method by visually comparing the generated and real images, as shown in Figures 6 and 7. HM translates the whole area of each image, and Reinhard method makes the contrast between local areas of the image more obvious; however, for the land survey dataset, there are obvious seasonal differences between the two temporal phases. The traditional method is not ideal for the reconstruction of bare soil to vegetation in the image, but the deep learning method can learn the deeper mapping relationships between images. DRIT migrates the color style of the target domain to the source domain but causes some building roof colors in the generated image to be close to the color of bare soil or vegetation. UNIT shares the latent space during translation, which can enhance the similarity between the local area of the source domain and the corresponding area of the target domain but ignores detailed information such as the edges of buildings. CycleGAN is more stable than the other methods on both datasets and can effectively reconstruct vegetation features and preserve the edge information of buildings.

To better analyze the effect of each method on the RGB bands of the images, we depicted the histogram information of the different generated images, as shown in Figures 8 and 9. HM can fit the data distribution of the target domain well, but it incurs a loss of semantic information due to the discontinuity of the digital number (DN) of the image. The Reinhard method makes the distribution of the DN more uniform and enhances the contrast of the image, but the whole distribution of the fitted data is poor. The wave peak of DRIT differs from the target domain, so it has an impact on the realism of the building roof color. The distribution of UNIT has a wider distribution range and enhances the contrast of the local area, but its wave response is not obvious. CycleGAN can fit the RGB distribution better than other methods.

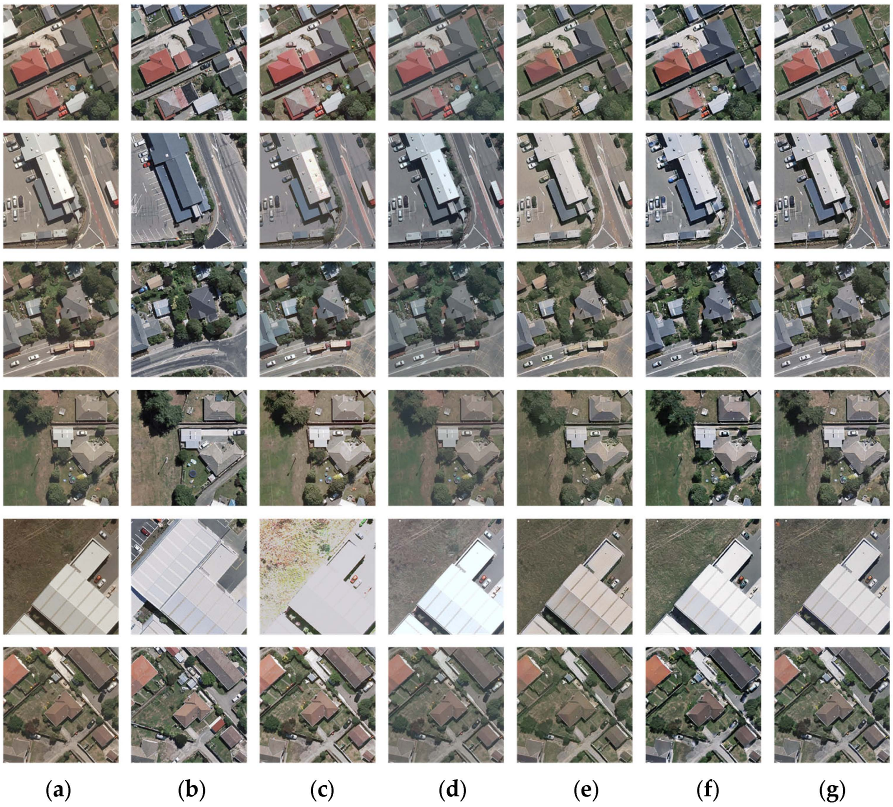


Figure 6. Image color translation results obtained on the Wuhan University building change detection dataset, (**a**) 2012 image, (**b**) 2016 image, (**c**) HM, (**d**) Reinhard method, (**e**) DRIT, (**f**) UNIT, and (**g**) CycleGAN.

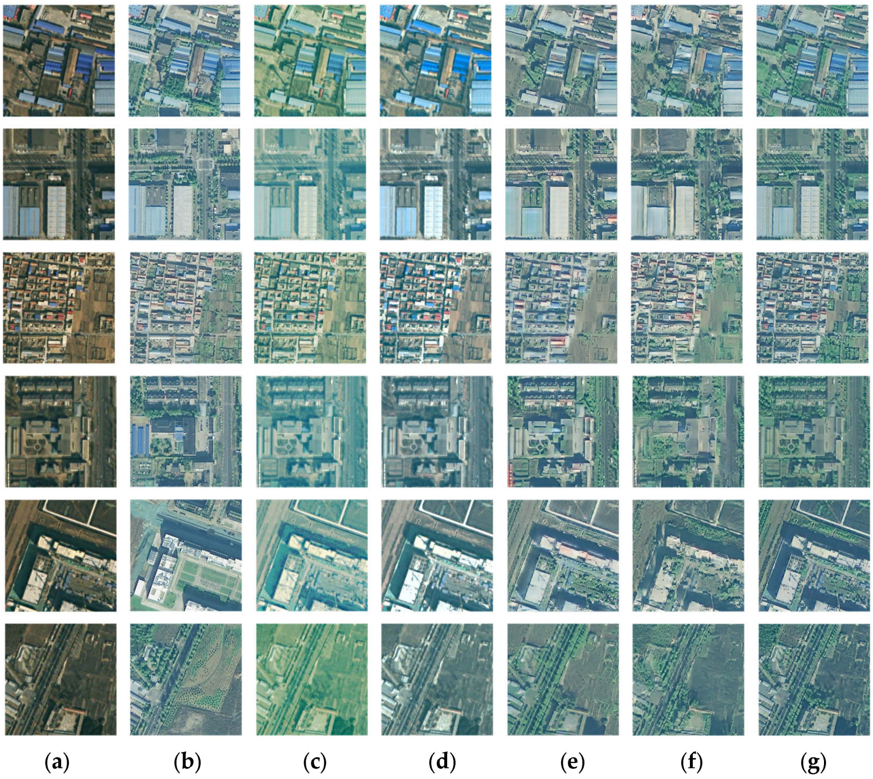


Figure 7. Image color translation results obtained on the Beijing Huairou district land survey dataset, (**a**) 2018 image, (**b**) 2019 image, (**c**) HM, (**d**) Reinhard method, (**e**) DRIT, (**f**) UNIT, and (**g**) CycleGAN.

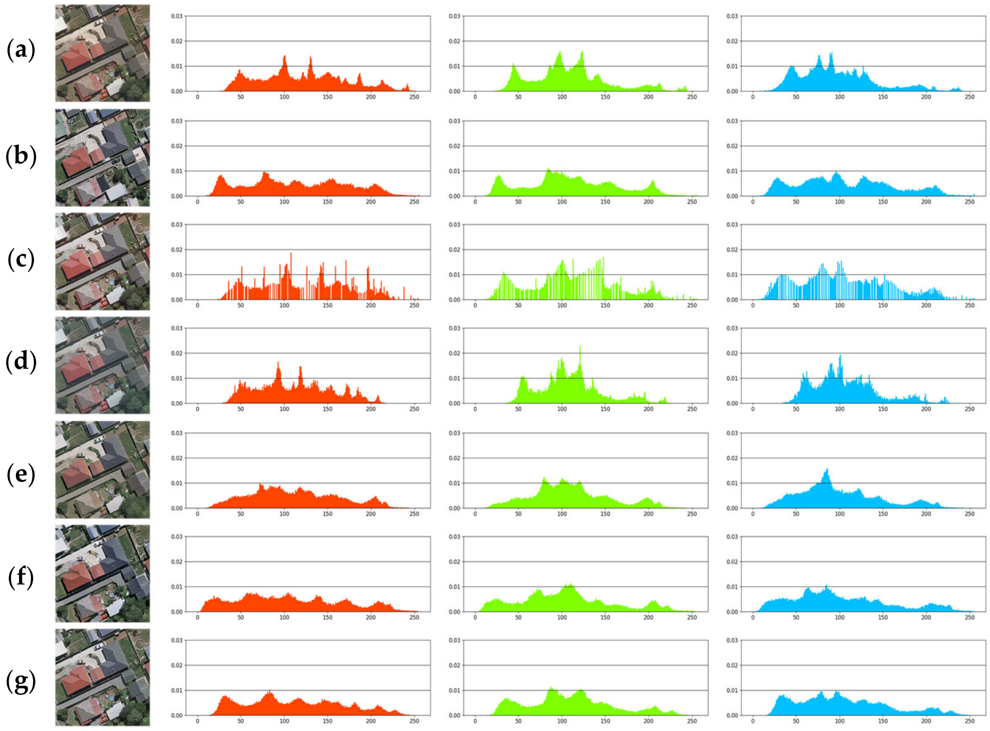


Figure 8. RGB histograms of the image color translation results obtained on the Wuhan University building change detection dataset, (**a**) 2012 image, (**b**) 2016 image, (**c**) HM, (**d**) Reinhard method, (**e**) DRIT, (**f**) UNIT, and (**g**) CycleGAN.

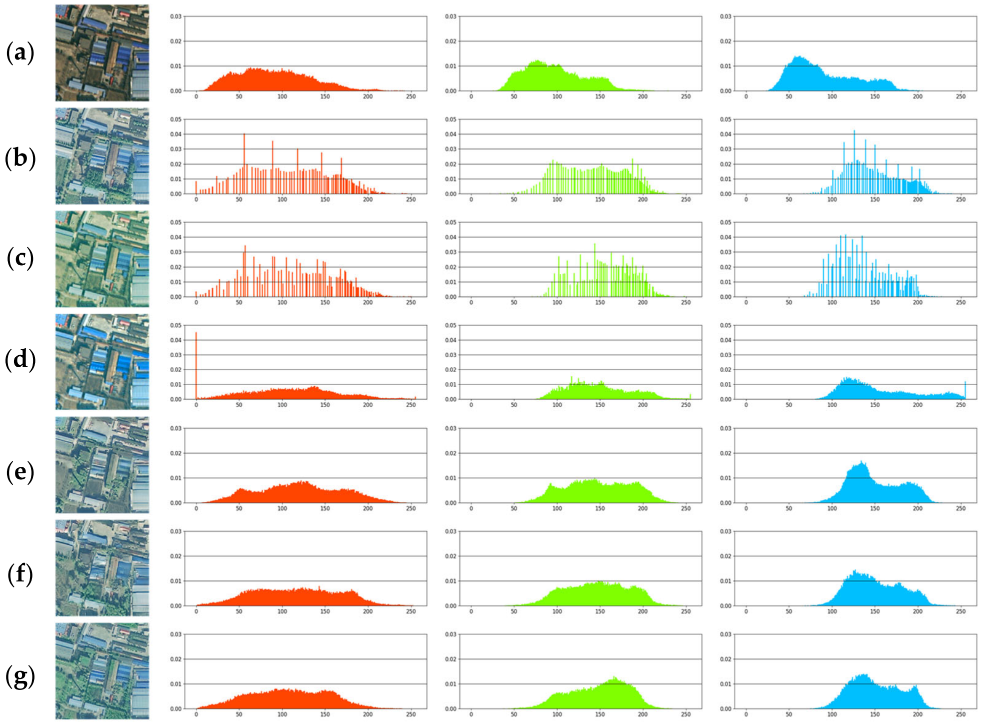


Figure 9. RGB histograms of the image color translation results obtained on the Beijing Huairou district land survey dataset, (**a**) 2018 image, (**b**) 2019 image, (**c**) HM, (**d**) Reinhard method, (**e**) DRIT, (**f**) UNIT, and (**g**) CycleGAN.

3.3. Numerical Results and Semantic Segmentation Visualization

We set the model with the original image without image color translation as the baseline and compared the gains achieved by other image color translation methods with the semantic segmentation model UNet(Eff-b1) in terms of five metrics.

The segmentation results obtained on the change detection dataset are shown in Table 1. Compared with the baseline, the use of image color translation can greatly improve the semantic segmentation performance for the buildings from the latest period. CycleGAN has the best overall performance compared with that of other methods, with IoU, precision, recall, accuracy, and F1 improvements of 10.93%, 9.22%, 2.94%, 2.43%, and 6.17%, respectively. HM achieves the optimal precision with a 9.62% improvement. In addition, for high-resolution images and datasets with small seasonal differences, there is little difference between the traditional and deep learning methods. Figure 10 shows the segmentation results of different methods. Holes and edge inaccuracies exist in the baseline results, in addition to certain degrees of false detections and missed detections, which are caused by the different data distributions of the two temporal images. The use of translation methods to align the data distributions can obtain better segmentation results, especially the CycleGAN method, which can help the segmentation model learn richer and more detailed information.

Table 1. Evaluation metrics of the semantic segmentation results obtained based on different image color translation methods for the Wuhan University building change detection dataset.

Method	IoU	Precision	Recall	Accuracy	F1
Baseline	0.8273	0.8775	0.9353	0.9635	0.9055
Histogram matching	0.9331	**0.9737**	0.9572	0.9871	0.9654
Reinhard method	0.9282	0.9671	0.9584	0.9861	0.9627
DRIT	0.9194	0.9601	0.9559	0.9843	0.9580
UNIT	0.9310	0.9665	0.9620	0.9867	0.9643
CycleGAN	**0.9366**	0.9697	**0.9647**	**0.9878**	**0.9672**

The bold values indicate the optimal values.

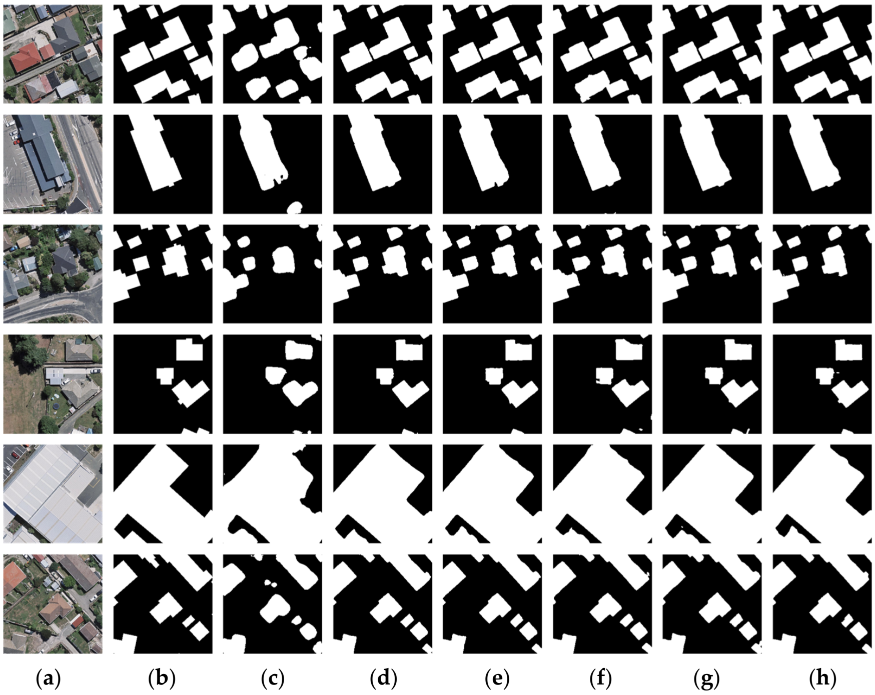


Figure 10. Segmentation results obtained by different image color translation methods on the Wuhan University building change detection dataset, (**a**) 2016 image, (**b**) 2016 ground truth, (**c**) Baseline, (**d**) HM, (**e**) Reinhard method, (**f**) DRIT, (**g**) UNIT, and (**h**) CycleGAN.

The segmentation results obtained on the land survey dataset are shown in Table 2. Due to the low resolutions and complex building labeling scenes in this dataset, training the model to predict the images of the latest period using only the pre-temporal data leads to a dramatic performance decrease. Compared with other methods, CycleGAN achieves the best results with 16.93%, 16.72%, 4.6%, 3.98%, and 11.38% improvements in the IoU, precision, recall, accuracy, and F1 metrics, respectively. It can be seen that the deep learning method outperforms the traditional methods in low-resolution and complex scenes and is able to learn deeper mapping relationships between different temporal images. The ground truths of the buildings in Figure 11 are divided using obvious roads or bare woodland. The baseline and traditional image color translation methods make the models unable to learn rich label information well, and the segmentation effects are poor. Deep learning-based translations can help the segmentation model better adapt to complex scenes.

Table 2. Evaluation metrics of the semantic segmentation results obtained by different image color translation methods on the Beijing Huairou district land survey dataset.

Method	IoU	Precision	Recall	Accuracy	F1
Baseline	0.6428	0.7134	0.8665	0.9291	0.7825
Histogram matching	0.6720	0.8093	0.7984	0.9426	0.8038
Reinhard method	0.7108	0.8370	0.8250	0.9506	0.8310
DRIT	0.7777	0.8705	0.8794	0.9630	0.8749
UNIT	0.8036	0.8789	0.9036	0.9675	0.8911
CycleGAN	**0.8121**	**0.8806**	**0.9125**	**0.9689**	**0.8963**

The bold values indicate the optimal values.

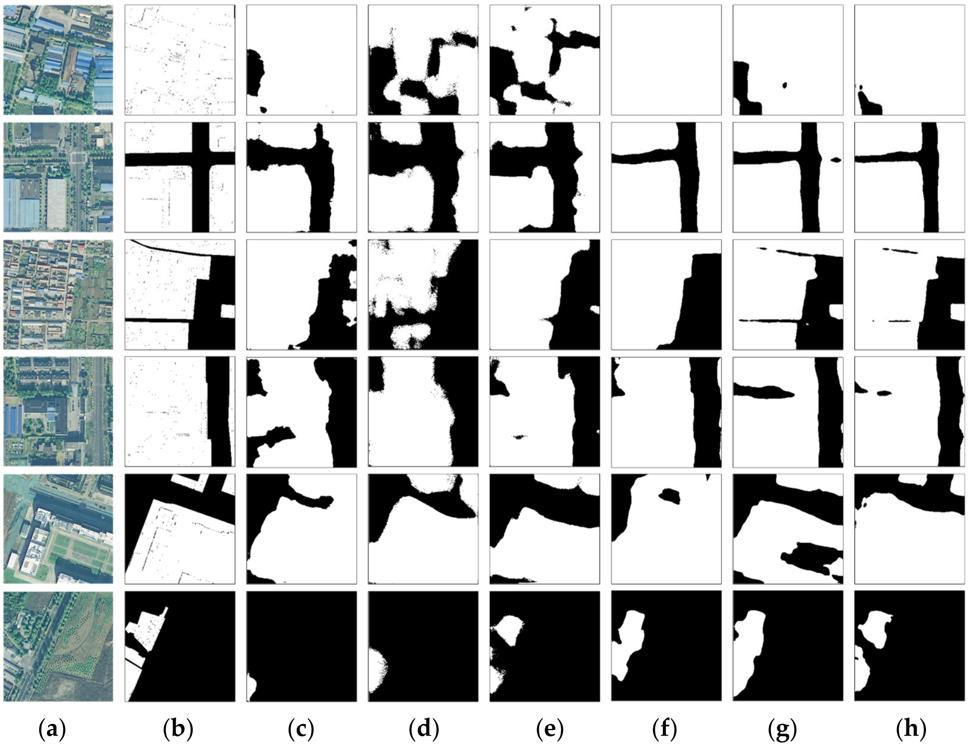


Figure 11. Segmentation results obtained by different image color translation methods on the Beijing Huairou district land survey dataset, (**a**) 2019 image, (**b**) 2019 ground truth, (**c**) baseline, (**d**) HM, (**e**) Reinhard method, (**f**) DRIT, (**g**) UNIT, and (**h**) CycleGAN.

In addition, we further investigated the effects of the images generated by CycleGAN on different semantic segmentation models, such as PSPNet, DeepLabV3, OCRNet, Segformer, SwinTransformer, UNet(ResNet50) and UNet(EfficientNet-b1). As seen from Tables 3 and 4, UNet with EfficientNet-b1 as the encoder achieves the best performance in terms of most of the metrics on the change detection dataset and the land survey dataset, outperforming the other competitive CNNs and transformer networks. On the change detection dataset, the IoU, precision, recall, accuracy, and F1 evaluation metrics of UNet(Eff-b1) segmentation are 0.9366, 0.9697, 0.9647, 0.9878, and 0.9672, respectively, while on the land survey dataset, the evaluation metrics are 0.8121, 0.8806, 0.9125, 0.9689, and 0.8963; the precision is 2.13% lower than that of DeepLabV3. In the case that the transformer results are lower than those of UNet(Eff-b1), we believe that the reasons for this are as follows, since the transformer captures global contextual information in an attentional manner to establish a long-distance dependence on the target object; however, the generated image after translation still has some distortion and distribution shifts compared with the real image, which causes errors to be accumulated several times when capturing the global context information and thus affects the final segmentation effect. In addition, there is more noise in low-spatial-resolution images, which further affects the application of the transformer network in remote sensing images. We also analyzed the efficiency levels of different models (Table 5). UNet using EfficientNet-b1 as an encoder is more efficient than ResNet50 and achieves better FLOPs 0.637(G) compared to most models, while its number of parameters is only 0.065(M) higher than that of the PSPNet. Compared to other models, UNet(Eff-b1) has a better balance between accuracy and complexity, so it can meet the needs of complex scenarios with large areas and the deployment of applications in real situations. From the prediction results of different models (Figures 12 and 13), the differences among the effects of different models on the change detection dataset are small, and the advantage of UNet(Eff-b1) is shown in the more accurate edges produced by small objects. For the land survey dataset, UNet(Eff-b1) has better visual effects; its results are not only globally closer to the ground truth but also have less jaggedness at the local edges. Therefore, UNet(Eff-b1) can achieve stable and accurate results under different labeling scenes and different resolutions.

Table 3. Evaluation metrics of the results obtained by different semantic segmentation methods on the Wuhan University building change detection dataset.

Method	IoU	Precision	Recall	Accuracy	F1
PSPNet	0.9153	0.9642	0.9474	0.9836	0.9557
DeepLabV3	0.9229	0.9685	0.9514	0.9851	0.9599
OCRNet	0.9301	0.9667	0.9609	0.9865	0.9638
Segformer	0.9133	0.9651	0.9445	0.9832	0.9547
SwinTransformer	0.9272	0.9629	0.9615	0.9859	0.9622
UNet(ResNet50)	0.9316	0.9655	0.9636	0.9868	0.9646
UNet(EfficientNet-b1)	**0.9366**	**0.9697**	**0.9647**	**0.9878**	**0.9672**

The bold values indicate the optimal values.

Table 4. Evaluation metrics of the results obtained by different semantic segmentation methods on the Beijing Huairou district land survey dataset.

Method	IoU	Precision	Recall	Accuracy	F1
PSPNet	0.7483	0.8535	0.8586	0.9575	0.8560
DeepLabV3	0.7678	**0.9019**	0.8377	0.9627	0.8686
OCRNet	0.7846	0.8780	0.8805	0.9644	0.8793
Segformer	0.7478	0.8420	0.8698	0.9568	0.8557
SwinTransformer	0.7814	0.7814	0.9025	0.9628	0.8773
UNet(ResNet50)	0.7637	0.8473	0.8856	0.9596	0.8660
UNet(EfficientNet-b1)	**0.8121**	0.8806	**0.9125**	**0.9689**	**0.8963**

The bold values indicate the optimal values.

Table 5. Efficiency levels of different semantic segmentation models.

Model **Backbone**	**PSPNet** **Res50** **(Depth = 3)**	**DeepLabV3** **Res50**	**OCRNet** **HR18**	**Segformer** **B2**	**SwinT** **S**	**UNet** **Res50**	**UNet** **Eff-b1**
Params (M)	**2.238**	39.634	12.026	2.478	48.746	32.521	2.303
FLOPs (G)	0.743	10.258	3.294	**0.381**	15.761	2.677	0.637

The bold values indicate the optimal values.

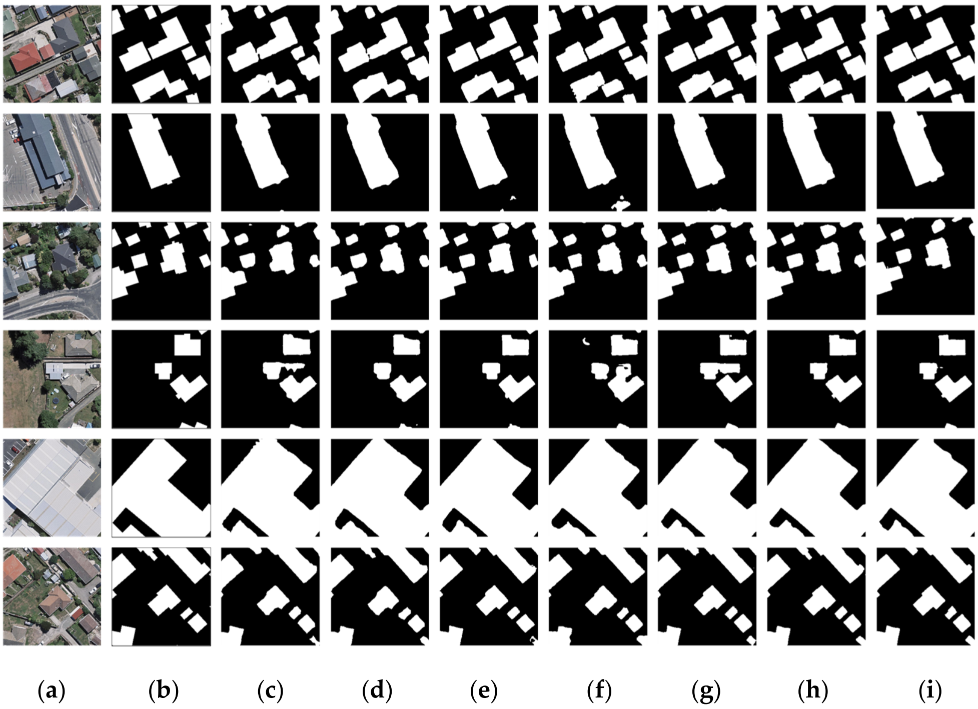


Figure 12. Results obtained by different semantic segmentation methods on the Wuhan University building change detection dataset, (**a**) 2016 image, (**b**) 2016 ground truth, (**c**) PSPNet, (**d**) DeepLabV3, (**e**) OCRNet, (**f**) Segformer, (**g**) SwinTransformer, (**h**) UNet(ResNet50), and (**i**) UNet(EfficientNet-b1).

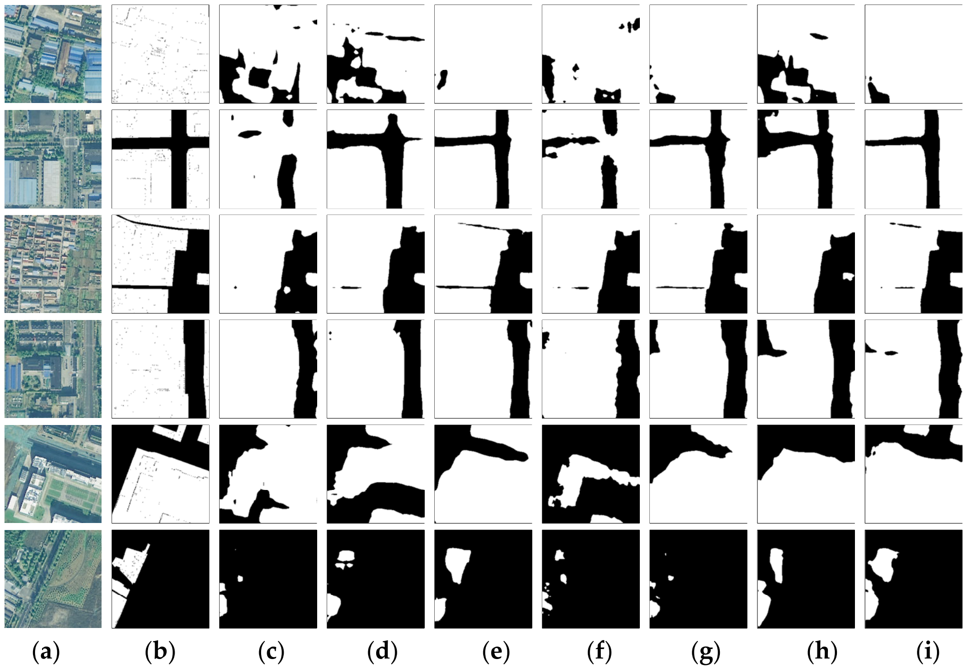


Figure 13. Results obtained by different semantic segmentation methods on the Beijing Huairou district land survey dataset, (**a**) 2019 image, (**b**) 2019 ground truth, (**c**) PSPNet, (**d**) DeepLabV3, (**e**) OCRNet, (**f**) Segformer, (**g**) SwinTransformer, (**h**) UNet(ResNet50), and (**i**) UNet(EfficientNet-b1).

3.4. Effectiveness Analysis of the Post-Processing Update Strategy

During the urban building update process, most of the buildings in the area remain unchanged, especially when the time interval between the acquisition of the two time-phased images is short. Moreover, the historical database is often already manually processed with strictly accurate labels for urban mapping, and these labels will be the primary reference for post-temporal ground truth. Therefore, we propose replacing some predictions that meet the overlap requirement with the corresponding historical labels to optimize the final update results. Tables 6 and 7 show the update results obtained for the two datasets using different thresholds after executing CycleGAN and UNet(Eff-b1). The update results are more dependent on the segmentation of the post-temporal images because of the long time interval between the pre- and post-temporal images in the change detection dataset and the higher image resolutions and more obvious changes in the local areas of the buildings. The post-processing update strategy works best when a threshold of 1 is used, i.e., it degenerates to a point where it uses only the predictions as the final update results. The reason for this is that there are certain coordinate shifts in some buildings in the pre- and post-temporal images, i.e., systematic errors are incurred in the topological position of the historical database to be replaced and the ground truths of the images from the latest period, which affects the evaluations of the update results and inhibits the effectiveness of the post-processing update strategy (Figure 14). It is worth noting that the updated results obtained with a threshold of 1 produce a small deviation from the accuracy above due to the post-processing of filtering non-polygon pixels and the effect of the simplified polygon of the contour extraction process.

Table 6. Update results obtained by post-processing update strategies with different thresholds on the Wuhan University building change detection dataset.

Threshold (θ)	IoU	Precision	Recall	Accuracy	F1
0	0.8490	0.9319	0.9051	0.9699	0.9183
0.2	0.8537	0.9359	0.9067	0.9709	0.9210
0.4	0.8555	0.9375	0.9072	0.9714	0.9221
0.6	0.8576	0.9396	0.9077	0.9718	0.9233
0.8	0.8622	0.9435	0.9091	0.9728	0.9260
1	**0.9363**	**0.9692**	**0.9649**	**0.9877**	**0.9671**

The bold values indicate the optimal values.

Table 7. Update results obtained by post-processing update strategies with different thresholds on the Beijing Huairou district land survey dataset.

Threshold (θ)	IoU	Precision	Recall	Accuracy	F1
0	0.9256	0.9477	**0.9754**	0.9884	0.9613
0.2	**0.9272**	0.9581	0.9663	**0.9888**	**0.9622**
0.4	0.9201	0.9609	0.9559	0.9877	0.9584
0.6	0.9104	**0.9618**	0.9445	0.9863	0.9531
0.8	0.8937	0.9603	0.9279	0.9837	0.9438
1	0.8120	0.8784	0.9148	0.9688	0.8962

The bold values indicate the optimal values.

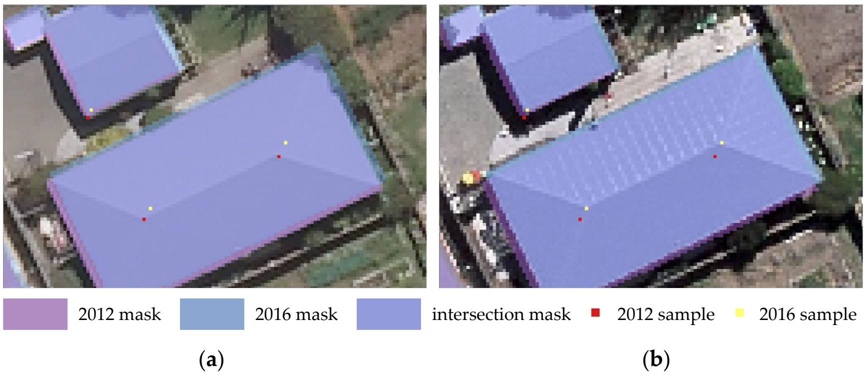


Figure 14. Offset of the Wuhan University building change detection dataset. (**a**) Red sample point and dull purple mask in 2012. (**b**) Yellow sample point and oxide blue mask in 2016. Light blue-purple represents intersection mask.

However, the update results are more dependent on the pre-temporal ground truth due to the shorter time interval between pre- and post-temporal images with lower image resolutions and wider label ranges. On the land survey dataset, the post-processing update strategy with a threshold of 0.2 greatly improves the update accuracy, with IoU, precision, recall, accuracy, and F1 metric improvements of 11.52%, 8.34%, 6.06%, 2%, and 6.6%, respectively. Moreover, the accuracy reached the best results at a threshold of 0.4 and the recall at a threshold of 0. As shown in Figure 15, using the post-processing update strategy to update the historical labels that satisfy the segmentation conditions can effectively optimize the update results by making full use of the a priori knowledge contained in the historical database. Column 5 predicts the added buildings in the changed area based on the accurate retention of the unchanged area, and column 6 updates the demolished buildings in the changed area without being influenced by the historical database.

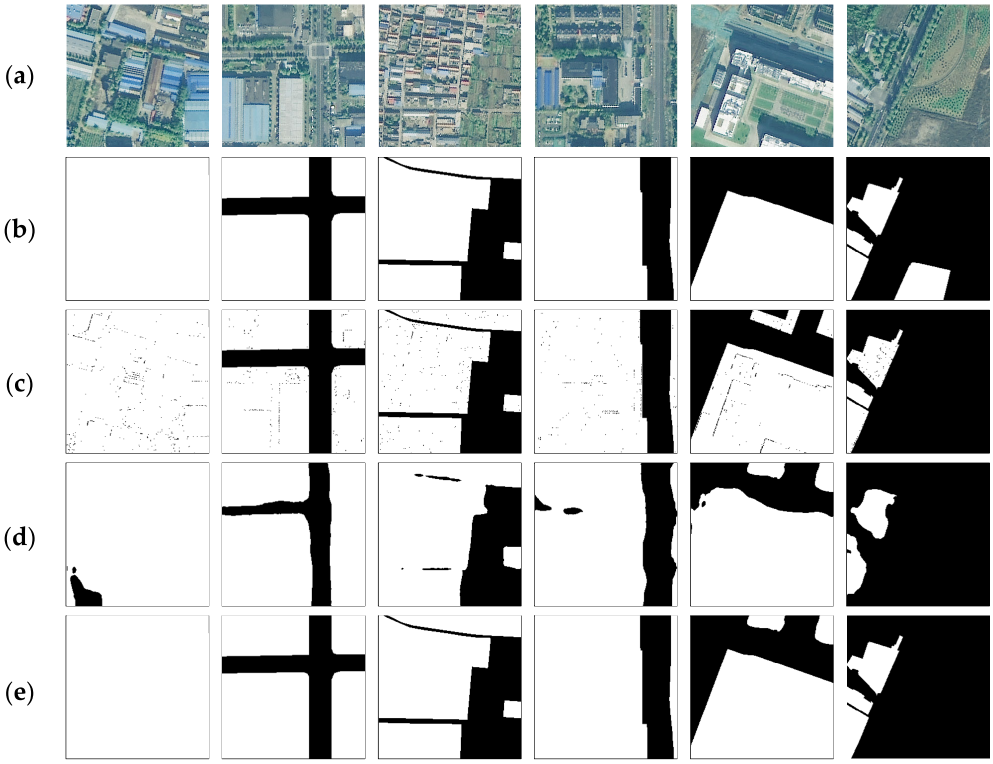


Figure 15. Update results obtained by the post-processing update strategies, (**a**) 2019 image, (**b**) 2018 ground truth, (**c**) 2019 ground truth, (**d**) CycleGAN, (**e**) CycleGAN-post-processing.

4. Discussion

Section 4.1 discusses the ablation experiments of the proposed update algorithm. Section 4.2 explains the reasons for the differences in thresholds across datasets.

4.1. Ablation Study

Ablation experiments of the proposed update algorithm were conducted on the change detection dataset and the land survey dataset. Here, UNet(Eff-b1) was trained on the original images without image color translation as a baseline, then CycleGAN and post-processing update strategy were gradually added on top of it to verify the effectiveness of each part of the update method. The results of the ablation experiments are shown in Tables 8 and 9.

Table 8. Ablation experiments of the proposed update algorithm on the Wuhan University building change detection dataset.

Methods	IoU	Precision	Recall	Accuracy	F1
Baseline	0.8273	0.8775	0.9353	0.9635	0.9055
Baseline + CycleGAN	0.9366	0.9697	0.9647	0.9878	0.9672
Baseline + CycleGAN + the Post-processing Update Strategy	0.9363	0.9692	0.9649	0.9877	0.9671

Table 9. Ablation experiments of the proposed update algorithm on the Beijing Huairou district land survey dataset.

Methods	IoU	Precision	Recall	Accuracy	F1
Baseline	0.6428	0.7134	0.8665	0.9291	0.7825
Baseline + CycleGAN	0.8121	0.8806	0.9125	0.9689	0.8963
Baseline + CycleGAN + the Post-processing Update Strategy	0.9272	0.9581	0.9663	0.9888	0.9622

After adding CycleGAN to the baseline, the IoU, precision, recall, accuracy, and F1 of the change detection dataset and the land survey dataset improved by 10.93%, 9.22%, 2.94%, 2.43%, 6.17%, and 16.93%, 16.72%, 4.6%, 3.98%, 11.38%, respectively, indicating that CycleGAN can mitigate the differences in the distribution of image color between different time phases. Figures 16 and 17 show the comparison of the visualization results from the first row to the third row, which show that the baseline can better recognize the edges of buildings in the images after adding CycleGAN, thus validating the effectiveness of CycleGAN.

After adding the post-processing update strategy to baseline + CycleGAN, the IoU, precision, recall, accuracy, and F1 of the land survey dataset improved by 11.52%, 8.34%, 6.06%, 2%, and 6.6%, respectively, suggesting that retaining strictly accurate historical labels is helpful for the improvement of the final update results. The change detection dataset, however, is subject to systematic errors due to coordinate shifts, so the segmentation results are used as the final update results.

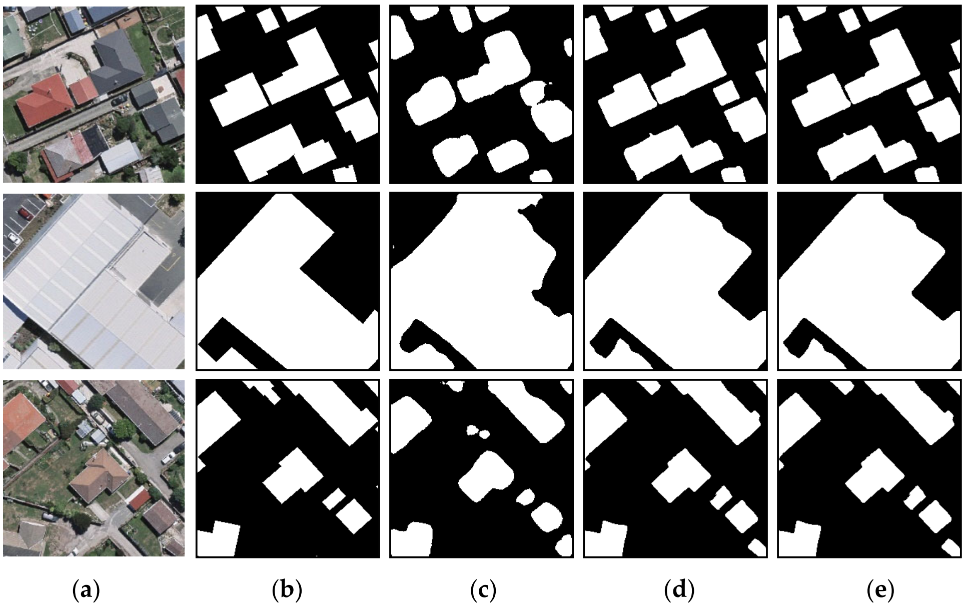


Figure 16. Comparison of ablation visualization results on the Wuhan University building change detection dataset, (**a**) 2016 image, (**b**) 2016 ground truth, (**c**) baseline, (**d**) baseline + CycleGAN, (**e**) baseline + CycleGAN + the post-processing update strategy.

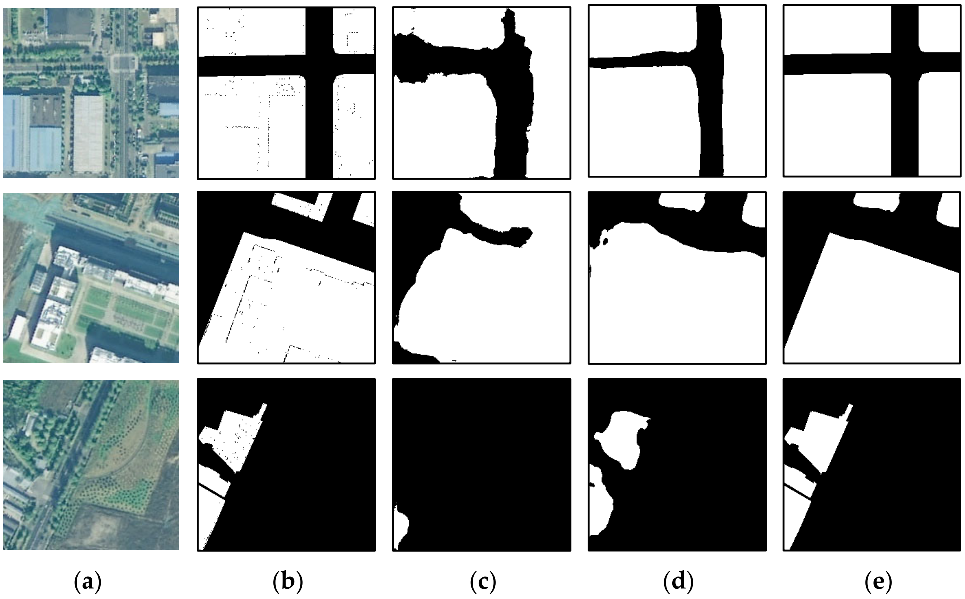


Figure 17. Comparison of ablation visualization results on the Beijing Huairou district land survey dataset, (**a**) 2019 image, (**b**) 2019 ground truth, (**c**) baseline, (**d**) baseline + CycleGAN, (**e**) baseline + CycleGAN + the post-processing update strategy.

4.2. Thresholds in the Post-Processing Update Strategy

The post-processing update strategy proposed in this paper was implemented based on intersection ratio, therefore, the intersection ratio of the pre-temporal label and the latter temporal segmentation result will affect the threshold of the post-processing update strategy. For the dataset with high image resolution, the semantic segmentation model can generally predict the target better, i.e., the main part of the building can be accurately predicted, and the error occurs more at the edges. Therefore, the high overlap between the segmentation result and the pre-temporal label leads to a larger threshold setting. However, for datasets with low image resolution or complex label range, it is extremely challenging for the semantic segmentation model to accurately predict the target, because the segmentation result is often part of the pre-temporal label or low overlap. Therefore, the partial overlap between the segmentation result and the previous temporal label leads to a smaller threshold setting.

In summary, different datasets need to set appropriate thresholds by considering the image resolution, label range, and the overlap between segmentation result and pre-temporal label. If the dataset is similar to the change detection dataset, the threshold can be set as large as possible. Conversely, if the dataset is similar to the land survey dataset, the threshold can be reduced appropriately.

5. Conclusions

In this paper, an update algorithm without manual relabeling is proposed to address the problem regarding differences between the data distributions of pre- and post-temporal images in the building update process. First, we used CycleGAN to reduce the color differences among satellite images under different time phases in an unsupervised way, then utilized UNet(Eff-b1) to learn the deep semantic information of buildings based on the generated images and historical database, and used this information to predict the images in the latest period. In addition, a post-processing update strategy is proposed to strictly retain the historical labels of unchanged regions. In an experiment, the characteristics of different image color translation methods, the improvements achieved by various semantic segmentation models and the effectiveness of post-processing update strategies were compared. The final IoU, precision, recall, accuracy, and F1 metrics of the update results obtained on the change detection dataset and land survey dataset are 0.9363, 0.9692, 0.9649, 0.9877, and 0.9671 and 0.9272, 0.9581, 0.9663, 0.9888, and 0.9622, respectively, which are improvements of 10.9%, 9.17%, 2.96%, 2.42%, and 6.16% and 28.44%, 24.47%, 9.98%, 5.97%, and 17.97%, respectively, over the baseline. However, this paper does not fully utilize the a priori knowledge contained in existing labels when using the image color translation method, and the post-processing update strategy needs to set appropriate thresholds according to different datasets. In future work, we will try to utilize the label information of the target category in the translation process to better couple it with the semantic segmentation model and study the characteristics of the changed and unchanged regions of different categories under multiple datasets to better utilize the label contours of the unchanged regions. In addition, we will further attempt to explore the applicability of the adaptive post-processing update strategy on the update task. The source code is publicly available at https://github.com/wangzehui20/building-footprints-update.

Author Contributions: Z.W. proposed the algorithm and wrote the whole article. Y.M. reviewed the whole paper carefully and gave some key suggestions. J.C. (Jingbo Chen) adjusted the structure of the article and corrected some mistakes. J.M. corrected some grammatical errors in the paper. A.Y. prepared and calibrated the data. J.C. (Jiansheng Chen) gave some suggestions on the format and writing style of the paper. 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 under grant number 2021YFB3900503 and the National Natural Science Foundation of China under grant number 61901471.

Data Availability Statement: The data presented in this study are available in this article.

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

References

1. Huang, X.; Cao, Y.; Li, J. An Automatic Change Detection Method for Monitoring Newly Constructed Building Areas Using Time-Series Multi-View High-Resolution Optical Satellite Images. *Remote Sens. Environ.* **2020**, *244*, 111802. [CrossRef]
2. 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]
3. Zheng, H.W.; Shen, G.Q.; Wang, H. A Review of Recent Studies on Sustainable Urban Renewal. *Habitat Int.* **2014**, *41*, 272–279. [CrossRef]
4. Cheng, J.; Mao, C.; Huang, Z.; Hong, J.; Liu, G. Implementation Strategies for Sustainable Renewal at the Neighborhood Level with the Goal of Reducing Carbon Emission. *Sustain. Cities Soc.* **2022**, *85*, 104047. [CrossRef]
5. Stiller, D.; Stark, T.; Wurm, M.; Dech, S.; Taubenböck, H. Large-Scale Building Extraction in Very High-Resolution Aerial Imagery Using Mask R-CNN. In Proceedings of the 2019 Joint Urban Remote Sensing Event (JURSE), Vannes, France, 22–24 May 2019; pp. 1–4.
6. Bouziani, M.; Goïta, K.; He, D.-C. Automatic Change Detection of Buildings in Urban Environment from Very High Spatial Resolution Images Using Existing Geodatabase and Prior Knowledge. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 143–153. [CrossRef]
7. Yang, G.; Zhang, Q.; Zhang, G. EANet: Edge-Aware Network for the Extraction of Buildings from Aerial Images. *Remote Sens.* **2020**, *12*, 2161. [CrossRef]
8. Liu, P.; Liu, X.; Liu, M.; Shi, Q.; Yang, J.; Xu, X.; Zhang, Y. Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network. *Remote Sens.* **2019**, *11*, 830. [CrossRef]
9. Zheng, J.; Tian, Y.; Yuan, C.; Yin, K.; Zhang, F.; Chen, F.; Chen, Q. MDESNet: Multitask Difference-Enhanced Siamese Network for Building Change Detection in High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3775. [CrossRef]
10. Deng, Y.; Chen, J.; Yi, S.; Yue, A.; Meng, Y.; Chen, J.; Zhang, Y. Feature Guided Multitask Change Detection Network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 9667–9679. [CrossRef]
11. Trenčanová, B.; Proença, V.; Bernardino, A. Development of Semantic Maps of Vegetation Cover from UAV Images to Support Planning and Management in Fine-Grained Fire-Prone Landscapes. *Remote Sens.* **2022**, *14*, 1262. [CrossRef]
12. Abubakar, F.M. Study of Image Segmentation Using Thresholding Technique on a Noisy Image. *Int. J. Sci. Res.* **2013**, *2*, 49–51.
13. Chakraborty, S. An Advanced Approach to Detect Edges of Digital Images for Image Segmentation. In *Applications of Advanced Machine Intelligence in Computer Vision and Object Recognition: Emerging Research and Opportunities*; IGI Global: Hershey, PA, USA, 2020; pp. 90–118.

14. Raja, N.; Fernandes, S.L.; Dey, N.; Satapathy, S.C.; Rajinikanth, V. Contrast Enhanced Medical MRI Evaluation Using Tsallis Entropy and Region Growing Segmentation. *J. Ambient. Intell. Humaniz. Comput.* **2018**, 1–12. [CrossRef]
15. Ke, L.; Xiong, Y.; Gang, W. Remote Sensing Image Classification Method Based on Superpixel Segmentation and Adaptive Weighting K-Means. In Proceedings of the 2015 International Conference on Virtual Reality and Visualization (ICVRV), Xiamen, China, 17–18 October 2015; pp. 40–45.
16. Bouman, C.A.; Shapiro, M. A Multiscale Random Field Model for Bayesian Image Segmentation. *IEEE Trans. Image Process.* **1994**, *3*, 162–177. [CrossRef] [PubMed]
17. Fan, J.; Yau, D.K.; Elmagarmid, A.K.; Aref, W.G. Automatic Image Segmentation by Integrating Color-Edge Extraction and Seeded Region Growing. *IEEE Trans. Image Process.* **2001**, *10*, 1454–1466. [PubMed]
18. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
20. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
21. Shang, R.; Zhang, J.; Jiao, L.; Li, Y.; Marturi, N.; Stolkin, R. Multi-Scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 872. [CrossRef]
22. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
26. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
27. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision Transformers for Remote Sensing Image Classification. *Remote Sens.* **2021**, *13*, 516. [CrossRef]
28. Zhang, J.; Zhao, H.; Li, J. TRS: Transformers for Remote Sensing Scene Classification. *Remote Sens.* **2021**, *13*, 4143. [CrossRef]
29. Lei, S.; Shi, Z.; Mo, W. Transformer-Based Multistage Enhancement for Remote Sensing Image Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–11. [CrossRef]
30. Li, X.; Xu, F.; Xia, R.; Li, T.; Chen, Z.; Wang, X.; Xu, Z.; Lyu, X. Encoding Contextual Information by Interlacing Transformer and Convolution for Remote Sensing Imagery Semantic Segmentation. *Remote Sens.* **2022**, *14*, 4065. [CrossRef]
31. Tasar, O.; Happy, S.L.; Tarabalka, Y.; Alliez, P. ColorMapGAN: Unsupervised Domain Adaptation for Semantic Segmentation Using Color Mapping Generative Adversarial Networks. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7178–7193. [CrossRef]
32. Yu, X.; Fan, J.; Zhang, M.; Liu, Q.; Li, Y.; Zhang, D.; Zhou, Y. Relative Radiation Correction Based on CycleGAN for Visual Perception Improvement in High-Resolution Remote Sensing Images. *IEEE Access* **2021**, *9*, 106627–106640. [CrossRef]
33. Zheng, Z.; Tang, X.; Yue, Q.; Bo, A.; Lin, Y. Color Difference Optimization Method for Multi-Source Remote Sensing Image Processing. *Proc. IOP Conf. Ser. Earth Environ. Sci.* **2020**, *474*, 042030. [CrossRef]
34. Yang, X.; Lo, C.P. Relative Radiometric Normalization Performance for Change Detection from Multi-Date Satellite Images. *Photogramm. Eng. Remote Sens.* **2000**, *66*, 967–980.
35. Reinhard, E.; Adhikhmin, M.; Gooch, B.; Shirley, P. Color Transfer between Images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 34–41. [CrossRef]
36. Schott, J.R.; Salvaggio, C.; Volchok, W.J. Radiometric Scene Normalization Using Pseudoinvariant Features. *Remote Sens. Environ.* **1988**, *26*, 1–16. [CrossRef]
37. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
38. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
39. Liu, M.-Y.; Tuzel, O. Coupled Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.
40. Liu, M.-Y.; Breuel, T.; Kautz, J. Unsupervised Image-to-Image Translation Networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
41. Huang, X.; Liu, M.-Y.; Belongie, S.; Kautz, J. Multimodal Unsupervised Image-to-Image Translation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 172–189.
42. Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; Yang, M.-H. Diverse Image-to-Image Translation via Disentangled Representations. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 35–51.

43. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to Discover Cross-Domain Relations with Generative Adversarial Networks. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1857–1865.
44. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
45. Alami Mejjati, Y.; Richardt, C.; Tompkin, J.; Cosker, D.; Kim, K.I. Unsupervised Attention-Guided Image-to-Image Translation. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
46. Xue, L.I.; Li, Z.; Qingdong, W.; Haibin, A.I. Multi-Temporal Remote Sensing Imagery Semantic Segmentation Color Consistency Adversarial Network. *Acta Geod. Cartogr. Sin.* **2020**, *49*, 1473.
47. Hoffman, J.; Tzeng, E.; Park, T.; Zhu, J.-Y.; Isola, P.; Saenko, K.; Efros, A.; Darrell, T. Cycada: Cycle-Consistent Adversarial Domain Adaptation. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 1989–1998.
48. Tan, M.; Le, Q. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
49. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2019**, *39*, 1856–1867. [CrossRef]
50. He, Y.R.; He, S.; Kandel, M.E.; Lee, Y.J.; Hu, C.; Sobh, N.; Anastasio, M.A.; Popescu, G. Cell Cycle Stage Classification Using Phase Imaging with Computational Specificity. *ACS Photonics* **2022**, *9*, 1264–1273. [CrossRef]
51. Le Duy Huynh, N.B. A U-Net++ with Pre-Trained Efficientnet Backbone for Segmentation of Diseases and Artifacts in Endoscopy Images and Videos. Available online: https://ceur-ws.org/Vol-2595/endoCV2020_paper_id_11.pdf (accessed on 14 October 2022).
52. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
53. Ji, S.; Wei, S.; Lu, M. 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. [CrossRef]
54. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv* **2016**, arXiv:1607.08022.
55. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.; Wang, Z.; Paul Smolley, S. Least Squares Generative Adversarial Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2794–2802.
56. Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; Webb, R. Learning from Simulated and Unsupervised Images through Adversarial Training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2107–2116.