remote sensing

MDPI

Article

Multi-SUAV Collaboration and Low-Altitude Remote Sensing Technology-Based Image Registration and Change Detection Network of Garbage Scattered Areas in Nature Reserves

Kai Yan [1,2], Yaxin Dong [1,2], Yang Yang [1,2,*] and Lin Xing [2,3]

[1] School of Information Science and Technology, Yunnan Normal University, Kunming 650500, China
[2] The Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University, Kunming 650500, China
[3] School of Physics and Electronic Information, Yunnan Normal University, Kunming 650500, China
* Correspondence: yangyang@ynnu.edu.cn or yyang_ynu@163.com

Abstract: Change detection is an important task in remote sensing image processing and analysis. However, due to position errors and wind interference, bi-temporal low-altitude remote sensing images collected by SUAVs often suffer from different viewing angles. The existing methods need to use an independent registration network for registration before change detection, which greatly reduces the integrity and speed of the task. In this work, we propose an end-to-end network architecture RegCD-Net to address change detection problems in the bi-temporal SUAVs' low-altitude remote sensing images. We utilize global and local correlations to generate an optical flow pyramid and realize image registration through layer-by-layer optical flow fields. Then we use a nested connection to combine the rich semantic information in deep layers of the network and the precise location information in the shallow layers and perform deep supervision through the combined attention module to finally achieve change detection in bi-temporal images. We apply this network to the task of change detection in the garbage-scattered areas of nature reserves and establish a related dataset. Experimental results show that our RegCD-Net outperforms several state-of-the-art CD methods with more precise change edge representation, relatively few parameters, fast speed, and better integration without additional registration networks.

Keywords: image registration with optical flow; end-to-end change detection; multi-SUAV low-altitude remote sensing

Citation: Yan, K.; Dong, Y.; Yang, Y.; Xing, L. Multi-SUAV Collaboration and Low-altitude Remote Sensing Technology-Based Image Registration and Change Detection Network of Garbage Scattered Areas in Nature Reserves. *Remote Sens.* **2022**, *14*, 6352. https://doi.org/10.3390/rs14246352

Academic Editor: Michalis Savelonas

Received: 12 November 2022
Accepted: 9 December 2022
Published: 15 December 2022

1. Introduction

Change detection (CD) is one of the major tasks in remote sensing (RS), which represents semantic changes in bi-temporal images by comparing satellite remote sensing images of the same region at different times and assigning binary labels to each pixel in the area. The technology of change detection using remote sensing images has been widely used in various fields and plays a crucial role in urban area expansion research [1,2], land use change analysis [3,4], forest vegetation cover monitoring [5–7] and natural disaster damage assessment [8,9]. In recent years, with the rapid development of small unmanned aerial vehicles (SUAVs) and the maturity of UAV low-altitude remote sensing technology, more and more researchers have focused on these. SUAV low-altitude remote sensing is convenient, in real-time and highly maneuverable. Many change detection studies use bi-temporal SUAV low-altitude remote sensing images to conduct experiments. However, due to the influence of wind changes and positioning system errors when the UAV is flying, there is a viewing angle change during imaging, which makes affine transformation between the bi-temporal images captured by SUAVs in the same area. The current studies often use additional registration methods to register bi-temporal images before performing

change detection, which not only adds additional research content but also makes the effect of registration directly affect the quality of change detection results.

Traditional registration methods apply feature descriptors to image pairs and use the nearest neighbor criterion to globally match keypoints to obtain pixel-to-pixel correspondences. However, these methods cannot accurately extract the correspondence between pairs of bi-temporal remote sensing images. This is because, in these image pairs, not only the change of the viewing point, but also the color change due to season, and the semantic change due to the ground object change are included. The traditional method cannot extract keypoints for matching by means of semantic information between bi-temporal remote sensing image pairs containing these changes. Owing to the amazing performance of convolutional neural networks (CNN) in extracting abstract semantic information in images, in recent years, many studies have used CNN to extract semantic features of bi-temporal remote sensing images, and perform registration by comparing pixels in feature space [10]. On the basis of CNN, recent optical flow methods exploit local correlation layers to estimate correspondences between image semantics, achieving great success in predicting semantic pixel-level accurate displacements. Nevertheless, the optical flow method evaluates the similarity of the local area around the pixel coordinates of the image. It is only suitable for small displacements, cannot capture large changes in viewing point and distance and performs poorly in SUAVs' low-altitude remote sensing image registration tasks.

Under the premise that the image pairs have been registered, for more and more complex semantic changes, the traditional pixel-based and object-based change detection methods are ineffective, and more and more researches focus on change detection using CNN. The network U-Net, which is designed for image segmentation, unexpectedly shows outstanding performance in change detection, establishing an encoder–decoder benchmark structure for subsequent research of change detection networks. Meanwhile, the method of extracting bi-temporal image features using a Siamese network [11–22] is widely used as a standard step in change detection. In order to improve the detection performance, some methods use a feature pyramid to extract multi-scale features during the down-sampling process of the encoder, which enriches the feature expression during up-sampling. Other methods utilize the attention mechanism during encoder down-sampling and decoder up-sampling to obtain better feature representation since attention-based methods (channel attention and spatial attention) are effective in establishing global information. Although the above methods have achieved good results in the change detection task, in these change detection networks using the encoder–decoder structure, continuous down-sampling and up-sampling will cause the loss of accurate location information in the shallow layers of the network, which will lead to the blurring in the edge of the change areas and the missed detection of small change areas in the detection results.

In this paper, we propose a convolutional neural network architecture, called RegCD-Net, for end-to-end change detection in the bi-temporal SUAV low-altitude remote sensing images. We build a multi-task CNN that implements bi-temporal low-altitude remote sensing image registration and subsequent change detection in a single network and it can be trained end-to-end. We use a combination of local and global correlation to solve the problem that bi-temporal low-altitude remote sensing image registration using the optical flow method performs poorly under large viewpoint changes and large displacement, and use the optical flow pyramid for layer-by-layer optimization. In order to solve the problem of loss of location information caused by continuous down-sampling in change detection, we use nested connections to combine deep semantic information and shallow location information and use the attention mechanism to perform network deep supervision and optimize feature representation.

The main contributions of this paper are as follows:

(1) We propose an end-to-end CNN architecture RegCD-Net, which integrates registration and change detection functions in a network, to achieve registration and change detection in bi-temporal SUAV low-altitude remote sensing images.

(2) We integrate global and local correlations to generate an optical flow pyramid and realize image registration through layer-by-layer optical flow fields.
(3) We utilize nested connections to combine effective information in different layers and perform deep supervision through a combined attention module to achieve change detection.
(4) We propose a method for generation change detection dataset with viewing angle changes using optical flow fields and generate a bi-temporal SUAV low-altitude remote sensing dataset for change detection in the garbage-scattered areas of nature reserves.

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 describes the proposed network architecture in detail. To evaluate our method, the experiments are designed in Section 4. The comparison experiment results are discussed in Section 5 and the ablution studies are shown in Section 6. Finally, the discussion and conclusions are presented in Sections 7 and 8.

2. Related Work

Change detection is an important task in computer vision and a critical method in remote sensing image analysis. When using this method to detect changes in bi-temporal low-altitude remote sensing image pairs, registration is needed for the image pairs first. In recent years, CNN has completely changed most fields in computer vision, and the method using CNN has been greatly improved compared with traditional methods. Here, we focus on CNN-based methods for image registration and change detection, as these are most relevant to our work.

The feasibility of using CNN for image registration originated from Spatial Transformer Network, a fully convolutional neural network built by Jaderberg et al. [23] for handwriting letter correction. However, the network structure is too simple, and it can only predict the deformation field of simple semantic images, which is not competent for realistic image registration. The motion relationship between bi-temporal images of the same scene in reality can also be regarded as the change (viewpoint change and semantic change) between the two images, and this change can be represented by an optical flow map. The optical flow map can describe the changing field between the two images, so that the two images can be registered by this field. Based on U-Net [24], Dosovitskiy et al. [25] proposed the first optical flow estimation network FlowNet, which directly estimated the optical flow between the original and target images by using the local correlation layer, providing strong clues for image registration. Ilg et al. [26] stacked several basic FlowNet models into a concatenated network FlowNet2, utilized correlation layers in a coarse-to-fine manner to estimate the similarity within the neighborhood interval of the center pixel, and used the intermediate optical flow to distort the target image for registration. Ranjan et al. [27] introduced an optical flow estimation network SpyNet combined with a spatial image pyramid model, which distorts the target image at each pyramid level through the current optical flow and calculates the updated optical flow to estimate the displacement between images layer by layer. The recent network PWC-Net proposed by Sun et al. [28] took the advantages of the above mentioned networks, combining spatial image pyramid, layer-by-layer distortion of intermediate optical flow and correlation cost volume in an optical estimation network, which is small and efficient.

Although the above-mentioned optical flow estimation models perform well when the images are in small deformations, they are not competent in obvious deformations, large displacements and significant differences in visual appearance (semantic changes) between images. With regard to this, Melekhov et al. [29] proposed DGC-Net, which exploits global correlation layers [30] to extract similarities between deep features and generates dense 2D correspondences to solve the optical flow prediction problem with strong geometric transformations between images. However, the network builds global cost volume through the coarsest resolution, which limits its accuracy in estimating small pixel displacement optical flow in high-resolution images.

In the field of change detection using CNN, the network U-Net proposed by Ronnerberger et al. [24] for image segmentation had shown outstanding performance in the change detection task due to its ability to extract deeper feature information, and established an encoder–decoder benchmark model structure for subsequent change detection research. Subsequently, Xiao et al. [31] and Guan et al. [32] applied the connection method of ResNet [33] and DenseNet [34] to U-Net, and proposed Res-Unet and FD-Unet respectively, which further enhanced the feature extraction ability of original U-net. Although the change detection models such as those based on U-net and its variants [18,35] can accurately predict the change areas, the encoding and decoding process of U-net is too direct and simple, and the accurate location information in shallow layers is often lost during successive down-sampling and up-sampling processes, leading to blurring in the change area edges and missed detection of small change areas.

In order to extract more accurate location features, some studies applied feature pyramid [15,17], such as STA-Net proposed by Bi et al. [36], which uses a feature pyramid to extract multi-scale features in the down-sampling process, enriching the representation of location information features during up-sampling. Other studies adopted attention mechanism [11–14,16,21], for example, Zhang et al. [37] proposed IFN-Net by adding the CBAM [38] to the change detection network framework and used CBAM to fuse features in the up-sampling process to enhance the accuracy of the boundary reconstruction of the detection results. Zhou et al. [39] proposed UNet++, which improves the region edge segmentation accuracy by adding nested dense connections and skip-path convolutional layers between the encoder and decoder, and adds a deep supervision mechanism to extract the output of different decoding layers. On the basis of this, Fang et al. [40] proposed SNUNet-CD, a change detection network based on Unet++. The network uses the Siamese network to extract bi-temporal image features and feeds these features into a densely connected encoder–decoder structure to output deep supervised features at different layers, and finally utilizes an attention mechanism to filter features for output. However, the network must input well-registered bi-temporal images, and images with poor registration effects will generate more non-semantic change areas through the network.

3. Materials and Methods

In this work, our goal is to detect the garbage-scattered areas in the nature reserves by means of change detection. We plan the flight paths of multiple SUAVs through the multi-UAV collaboration platform, realize the capture of bi-temporal images of the large-scale ground in the nature reserve and use the CNN to accomplish the detection of changes in the garbage-scattered areas in the bi-temporal images.

In the classic remote sensing image change detection task using CNN, the bi-temporal images usually need to be pre-registered to ensure that the perspectives of the two images are consistent, and the change detection can be performed. Among them, registration and change detection are two separate processes that need to be accomplished by using different networks. Different from these approaches that separate the registration process from change detection, in this work, we treat registration and change detection as two consecutive tasks in the same network, and proposed an end-to-end network to address the registration pretreatment of bi-temporal images and subsequent change detection issues.

The architecture of the proposed network is presented in Section 3.1. In Sections 3.2 and 3.3, the optical flow registration subnetwork and subsequent change detection subnetwork of the overall network architecture are described in detail. Section 3.4 provides details of the multi-SUAV collaboration platform. In Section 3.5, the training details are presented in detail, including the structure of the loss function and the generation of the dataset.

3.1. Network Architecture

The proposed network is an encoder–decoder structure overall. As shown in Figure 1, the network is mainly composed of three parts: down-sampling backbone, optical flow reg-

istration sub-network and nested connection change detection subnetwork. The bi-temporal images are firstly encoded by a down-sampling backbone, and then the two images are registered through the optical flow registration subnetwork. Finally, the registered pair of images is sent to the nested connection change detection subnetwork for decoding to obtain the final change detection result.

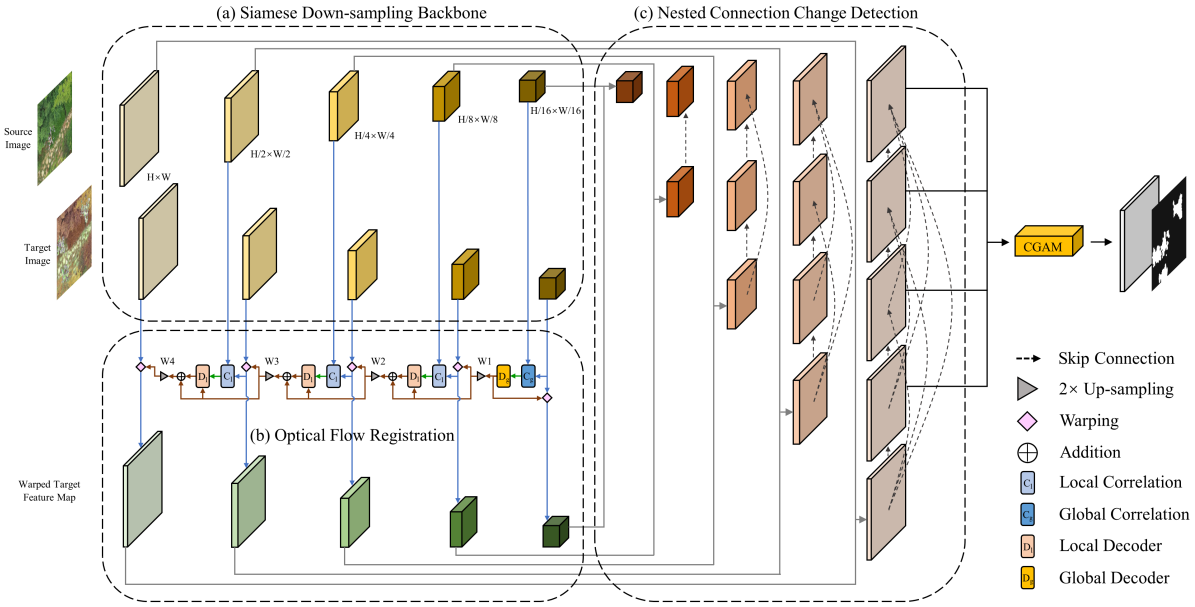

Figure 1. The architecture of the RegCD network.

The ResNet18 [33] backbone is used as the down-sampling network in the encoder. Furthermore, the siamese down-sampling network, which is constituted by two parallel ResNet18 backbones, is designed for simultaneous down-sampling of source and target images. As shown in Figure 1a, bi-temporal images are input into two branches of the siamese down-sampling network and down-sampled simultaneously. Owing to weight sharing between the two branches, two down-sampling networks with the same set of parameters can extract approximately the same feature in two images. Then, the concatenation is used to fuse the two extracted feature maps into a single one that contains the same and different features of the two maps. It should be emphasized that one of the feature maps involved in fusion is the registered feature map obtained after warping by optical flow.

Before the decoder performs up-sampling, the input bi-temporal images are necessary to register to ensure that two images are in the same perspective. That is to say, the feature maps of target images in each level are registered to the perspective where the corresponding feature maps of the source images are located. The estimated displacement field $\mathbf{w} \in \mathbb{R}^{H \times W \times 2}$, which is often called optical flow, is used to warp the target images I_t to the source images I_s as follows:

$$I_s(\mathbf{c}) = I_t(\mathbf{c} + \mathbf{w}(\mathbf{c})) \tag{1}$$

where $\mathbf{c} \in \mathbb{Z}^2$ is the coordinate of each pixel in images. Through field $\mathbf{w}$, the coordinate $\mathbf{c}$ in the target images can be mapped directly to its corresponding location in the source images to complete the registration of the two images.

The optical registration is applied to each feature map level and accomplished layer by layer from deep to shallow layers of the network. As shown in Figure 1b, in the deepest layer of the down-sampling, the two feature maps are sent to the global correlation module *GC* to calculate global cost volume, and then the result is sent to the global mapping decoder *GD* to estimate an optical flow field. The target feature map in the deepest layer is warped through the optical flow field, and the field will be sent to the previous layer

after 2× up-sampling. In the shallow layers, the target feature map is firstly warped by an optical flow field from the deeper layer, and then the local cost volume is calculated together with the source feature map in the local correlation module *LC*. The optical flow field from the deeper layer is not only decoded by the local mapping decoder *LD* together with the local cost volume into a new optical flow field, but also added to this new field to finally generate the optical flow field of this layer. The optical flow filed **w** in shallow layers is defined as:

$$\mathbf{w}^l = LD\left[LC\left(Warp\left(F_t^l, U(\mathbf{w}^{l-1})\right), F_s^l\right), U(\mathbf{w}^{l-1})\right] + U(\mathbf{w}^{l-1}) \tag{2}$$

where U denotes the 2× up-sampling, and l denotes the level of the feature layer. F_s and F_t refer to the feature maps of the source and target images. LC is the operation of computing the local correlation, and LD is the decoding operation using a local mapping decoder. $Warp$ is the operation that uses the optical flow field to warp the target feature map to achieve registration. In the top layer, the target image will be directly warped to generate the registered original-size image. More details about optical registration are available in Section 3.2.

The nested connection up-sampling structure acts as a decoder to densely up-sample the feature map of each layer. The paired source and warped target feature maps of each layer in down-sampling are sent to the corresponding feature map size up-sampling network together. The semantic and spatial information in the feature maps of different layers are fused through densely nested connections, and the four feature maps of original size are generated by stepwise up-sampling. Eventually, these four feature maps are convolved through an attention module to generate a final change detection map, as shown in Figure 1c. Please refer to Section 3.3 for details of the nested connection and change detection. The integral inference detail of our RegCD-Net is shown in Algorithm 1.

Algorithm 1 Inference of RegCD-Net for change detection

Input: $\mathbf{I} = \{(\mathbf{I}_s, \mathbf{I}_t)\}$ (a pair of unregistrated images)
Output: C (a prediction change mask)

// step1: extract multi-layer features by a siamese down-sampling backbone
for *i in* $\{s, t\}$ **do**
  $\mathbf{F}_i$ = Down-sampling_Backbone($\mathbf{I}_i$)
end
// step2: use optical flow field to warp target feature maps for registration
if *layer* $l = 1$ **then**
  $\mathbf{W}^1$ = Global_Decoder(Global_Corrleation($\mathbf{F}_s^1$, $\mathbf{F}_t^1$))
  $\widetilde{\mathbf{F}}_t^1$ = Warp($\mathbf{F}_t^1$, $\mathbf{W}^1$)
else
  $\widetilde{\mathbf{F}}_t^l$ = Warp($\mathbf{F}_t^l$, Up-sampling($\mathbf{W}^{l-1}$))
  $\Delta\mathbf{W}^l$ = Local_Decoder(Local_Correlation($\mathbf{F}_s^l$, $\widetilde{\mathbf{F}}_t^l$), Up-sampling($\mathbf{W}^{l-1}$))
  $\mathbf{W}^l$ = $\Delta\mathbf{W}^l$ + Up-sampling($\mathbf{W}^{l-1}$)
end
// step3: use nested connection to up-sample feature maps for change detection
for *i in* $\{0, 1, 2, 3, 4\}$ **do**
  if $j = 0$ **then**
    $\mathbf{X}^{i,0}$ = Conv(Concatenate($\mathbf{F}_s^i$, $\widetilde{\mathbf{F}}_t^i$)))
  else
    $\mathbf{X}^{i,j}$ = Conv(Concatenate($[\mathbf{X}^{i,k}]_{k=0}^{j-1}$, Up-sampling($\mathbf{X}^{i+1,j-1}$))
  end
end
// step4: obtain change mask by attention module
C = Conv(CGAM($\mathbf{X}^{0,1}$, $\mathbf{X}^{0,2}$, $\mathbf{X}^{0,3}$, $\mathbf{X}^{0,4}$))

3.2. Optical Flow Registration

The optical flow is generated by a sequence of layers in the optical flow estimate network to measure local and global correspondence between the source and target feature images in each layer. The similarity calculated results of two feature maps, generally called cost volume [25], can quantify the correspondence between two feature maps in each layer and provide a strong basis for the network to estimate optical flow layer by layer. According to the calculation range of the feature correspondence relationship, the cost volume can be calculated in the way of local correlation and global correlation.

3.2.1. Local Correlation

The local correlation layer only calculates the feature correspondence in the neighborhood pixel between the source and target feature map [41,42]. The local correlation LC^l between the source $F_s^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ and target $F_t^l \in \mathbb{R}^{H_l \times W_l \times C_l}$ feature maps is defined as:

$$LC^l(\mathbf{c}, \mathbf{o}) = F_t^l(\mathbf{c})^T F_s^l(\mathbf{c} + \mathbf{o}), \quad \|\mathbf{o}\|_\infty \leq R \tag{3}$$

where l refers the the level in the feature pyramid layer, $\mathbf{c} \in \mathbb{Z}^2$ is a coordinate in the target feature map and $\mathbf{o} \in \mathbb{Z}^2$ is the offset from this coordinate. The maximum offset in any direction is always constrained in the neighborhood radius R, and the correlations are only calculated in this neighborhood radius. In theory, the correlation result is a 4D tensor, which is positions offset combine of two 2D tensor. In practice, the cost volume LC^l is organized as a 3D tensor, which has the size of $H^l \times W^l \times (2R+1)$.

3.2.2. Global Correlation

The global correlation layer is only used in the deepest feature layer to calculate the global correlation between the most roughly feature maps. It evaluates the correspondence in all locations between the source and target feature maps [43,44]. The global correlation GC^l is defined as follows:

$$GC^l(\mathbf{c}, \mathbf{c}') = F_t^l(\mathbf{c})^T F_s^l(\mathbf{c}') \tag{4}$$

where $F_s^l(\mathbf{c}')$ and $F_t^l(\mathbf{c})$ refer to the feature map extracted from all source feature map coordinates and target feature map coordinates, respectively. The cost volume GC^l is organized as a 3D tensor of size $H^l \times W^l \times (H^l W^l)$.

3.2.3. Local and Global Correlation Assemble

According to their range of correlation calculation, the behaviors of the local and global correlation layers present some complementary characteristics. The local correlation layer is widely used in the optical flow estimate network to evaluate the displacements of two feature maps. Limited by neighborhood radius calculation, the local correlation layer can be applied in high-resolution feature maps to estimate small displacements precisely. That is to say, the correlation calculations are restricted to a small range by local correlation, and failed to estimate a large offset of source and target feature maps. In contrast, the global correlation calculates a large range correspondence relationship without maximum range radius limitation; therefore, it can estimate the large-scale displacement of the two feature maps.

Moreover, the cost volume of the global correlation is calculated as a tensor in the size of $H^l \times W^l \times (H^l W^l)$, which refers to a coordinate in the source feature maps that need to calculate two direction offsets with all coordinates in target feature maps. In the high-resolution feature maps, the very large space complexity $O((H^l W^l)^2)$ will cause the memory to be occupied by a huge amount of tensor computation. Hence, the global correlation layer is only utilized in calculating the most coarse-resolution feature maps. The architecture has a combination of local and global correlation, which is used in the proposed network to estimate optical flow (Figure 2).

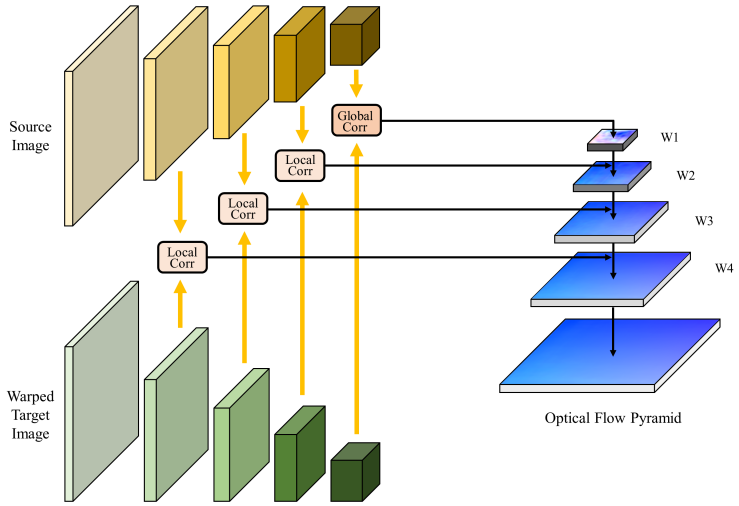

Figure 2. The architecture of the optical flow pyramid with global and local correlations.

3.2.4. Flow Decoder

The flow decoder is used to estimate the optical flow field of this layer through the optical flow field generated by the previous layer and the cost volume produced by the correlation of this layer. In the deepest feature layer, only the resulting global correlation GC needs to be sent into the flow decoder GD, as implemented in DGC-Net [29], to estimate the optical flow field $\mathbf{w}^1$ at the coarsest level of the feature layers:

$$\mathbf{w}^1 = GD\Big(GC(F_t^1, F_s^1)\Big) \tag{5}$$

In the rest level l of the feature layers, the residual flow $\Delta\mathbf{w}^l$, which is the computing result of the flow decoder LD, is defined as:

$$\Delta\mathbf{w}^l = LD\Big(LC(\widetilde{F}_t^l, F_s^l; R), U(\mathbf{w}^{l-1})\Big) \tag{6}$$

where LC refers the local correlation with search radius R, and $U(\cdot)$ denotes the $2\times$ up-sampling. $\widetilde{F}_t^l$ is the registered target feature map F_s^l warped by the up-sampled optical flow field $U(\mathbf{w}^{l-1})$ from the deeper one layer, which is defined as:

$$\widetilde{F}_t^l(\mathbf{c}) = F_t^l\Big(\mathbf{c} + U\Big(\mathbf{w}^{l-1}(\mathbf{c})\Big)\Big) \tag{7}$$

where $\mathbf{c}$ is a coordinate in the maps. The complete optical flow field in layer l is defined as:

$$\mathbf{w}^l = \Delta\mathbf{w}^l + U(\mathbf{w}^{l-1}) \tag{8}$$

The flow decoder GD and LD both consist of five convolutional layers with dense connections [34]. The number of channels in each convolutional layer is 128, 128, 96, 64 and 32, respectively, and the size of all convolutional kernels is 3×3. Finally, the estimated optical flow field is output through a 2D linear convolution.

3.2.5. Optical Flow Pyramid

In a typical CNN network structure, as the down-sampling process continues to deepen, the feature maps in deeper layers contain more rich semantic information. However, accurate location information in shallow layers is continuously lost in the down-sampling process. The feature pyramid structure [15,27,28,45] is used to combine the rich semantic information in deeper feature maps and the accurate location information in shallow feature maps to achieve more precise feature representation.

Similarly, the feature pyramid structure in this paper is utilized for fusion of optical flows from different feature layers. The optical flow in coarse layers can estimate large range displacement, while the optical flow in high-resolution layers is used to determine tiny offsets precisely. During the optical flow up-sampling process, the optical flow pyramid fuses the deep and shallow optical flows layer by layer, eventually generating an optical flow that can estimate the exact displacement of the two images at the original resolution. The optical flow pyramid architecture is shown in Figure 2.

3.3. Nested Connection Change Detection

The standard encoder–decoder architecture is widely used in change detection. The deeper the feature extraction layer of an encoder, the richer the semantic information of the feature map that will be extracted, while the extracted location information will be more vague. To tackle this contradiction, the decoder structure is added to fuse the semantic information of the deep layer with the location information shallow layer, and up-sampling to the original resolution step by step. Sparked by the ResNet, in order to achieve better fusion information, the skip connection is used in the decoder to connect more deep and shallow layers when up-sampling.

3.3.1. Nested Connection Up-Sampling

Different from the traditional network of the encoder–decoder structure, which connect encoder and decoder feature maps straightly, in UNet++ [39], this direct connection is expanded as dense connections by the use of skip connection. In order to maintain location information in shallow feature maps and semantic information in deep ones, to bridge the semantic gap between the encoder and decoder feature maps, the dense skip connection between the encoder and decoder is used in the proposed network.

The dense skip connection is shown in Figure 3. The source and target feature maps are down-sampled by two branched feature extraction backbones. Then, in each feature layer level, the extracted feature maps of two branches are concatenated to generate a single one that contains the same and different features of the two maps. The concatenated feature maps participate in the nested connection up-sampling process, transmitting the feature information from different layers to the decoder through skip connections, and compensating for the loss of location information in deep layers. For example, the feature maps of ${X_s}^{2,0}$ and ${X_t}^{2,0}$ are extracted by two branches of down-sampling backbone, then the $X^{2,0}$ is generated by convolution after concatenating the two. The three block of ${X_s}^{3,0}$, ${X_t}^{3,0}$ and $X^{3,0}$ are the one-level deeper counterparts, respectively, of ${X_s}^{2,0}$, ${X_t}^{2,0}$ and $X^{2,0}$. To obtain $X^{2,2}$, $X^{3,1}$ is obtained firstly by convolution after concatenating $X^{3,0}$ with $2\times$ up-sampled $X^{4,0}$. Then, the intermediate unit $X^{2,1}$ is generated by convolving the concatenation of $X^{2,0}$ and $2\times$ up-sampled $X^{3,0}$. At last, through skip connection, $X^{2,0}$ could be concatenated with $X^{2,1}$ and $X^{3,1}$, eventually generating $X^{2,2}$ by convolution. Up-sampling is needed for every unit except for the original size unit, in order to achieve dense nested connections throughout the up-sampling process of the decoder.

The convolution block is designed as a residual unit structure [33] through a skip connection. Each convolution block is preceded by a concatenation block which concatenates the output of the previous convolution block in the same feature map level with the up-sampled output of the one-level deeper convolution block and unified by convolution blocks. The structure of the convolution block is shown in Figure 4.

Formally, let $x^{i,j}$ denote the output of unit $X^{i,j}$, where i denotes the down-sampling layer level and j denotes the number of skip connections received by this unit. The $x^{i,j}$ is defined as follows:

$$x^{i,j} = \begin{cases} C\left(\left[x_s^{i,j}, x_t^{i,j}\right]\right), & j = 0 \\ C\left(\left[\left[x^{i,k}\right]_{k=0}^{j-1}, U(x^{i+1,j-1})\right]\right), & j > 0 \end{cases} \tag{9}$$

where function $C(\cdot)$ denotes the convolution operation of the convolution block, function $U(\cdot)$ is a 2×up-sampling operation, and $[\cdot]$ denotes the concatenation operation on the channel dimension. Specifically, $x_s^{i,0}$ and $x_t^{i,0}$ are sourced from the down-sampling, and other units $x^{i,j}$ at level $j = 0$ are concatenated by $x_s^{i,0}$ and $x_t^{i,0}$; units at level $j > 0$ receive both the outputs of the previous units in the same sampling level and an up-sampled output of the deeper unit. The location and semantic information in different levels of encoders are transmitted to the decoders in succession for concatenation, convolution and up-sampling through these dense skip connections.

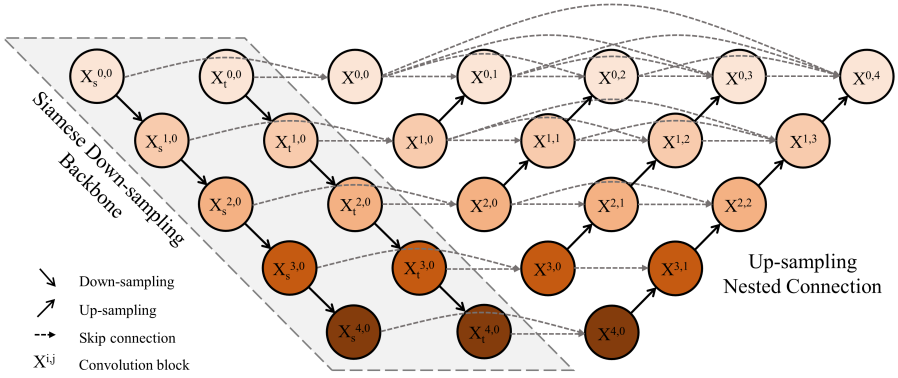

Figure 3. The architecture of up-sampling nested connection.

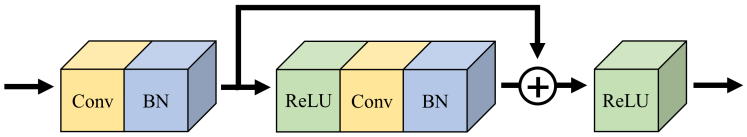

Figure 4. The structure of the convolution block.

3.3.2. Channel Attention

The outputs of the nested connection up-sampling are four feature maps, which are the outputs of the unit $x^{0,1}$, $x^{0,2}$, $x^{0,3}$ and $x^{0,4}$, with the same size as the original images. However, the four outputs have different representations of semantic levels and spatial location, because they are generated through different levels of skip connection and up-sampling pathways. The outputs from shallow pathways have precise location information and finer-grained features, by contrast, the outputs from deep pathways have richer semantic information and coarse-grained features. Therefore, a select mechanism is needed to screen out effective feature information representation when fusing the four feature maps.

On the basis of the channel attention module (CAM) [38], a select mechanism called channel group attention module (CGAM) is proposed in this paper to select more appropriate feature information and focus on more effective feature representations between each set of feature maps. As shown in Figure 5, the four groups of output feature maps are concatenated first, and then a CAM is used to extract the inter-group channel relationship. Meanwhile, another CAM is also used to extract intra-group relationships after summing the four groups of feature maps. Finally, the final refined output is obtained by sequentially multiplying the concatenated feature map and the two CAMs. In short, the CGAM of feature map $F \in \mathbb{R}^{H \times W \times C}$ is defined as follows:

$$CAM(F) = \sigma(MLP(AP(F)) + MLP(MP(F))) \tag{10}$$

$$M_{intra} = CAM(x^{0,1} + x^{0,2} + x^{0,3} + x^{0,4}) \tag{11}$$

$$M_{inter} = CAM\Big([x^{0,1}, x^{0,2}, x^{0,3}, x^{0,4}]\Big) \tag{12}$$

$$CGAM = M_{inter} \otimes \left[M_{intra} \otimes x^{0,1}, M_{intra} \otimes x^{0,2}, M_{intra} \otimes x^{0,3}, M_{intra} \otimes x^{0,4} \right] \quad (13)$$

where σ denotes the sigmoid function, and MLP is the multi-layer perception layer. AP and MP denote average pooling and max pooling operations, respectively. $[\cdot]$ denotes the concatenation operation, and $\otimes$ denotes the element-wise multiplication between feature maps and attention maps.

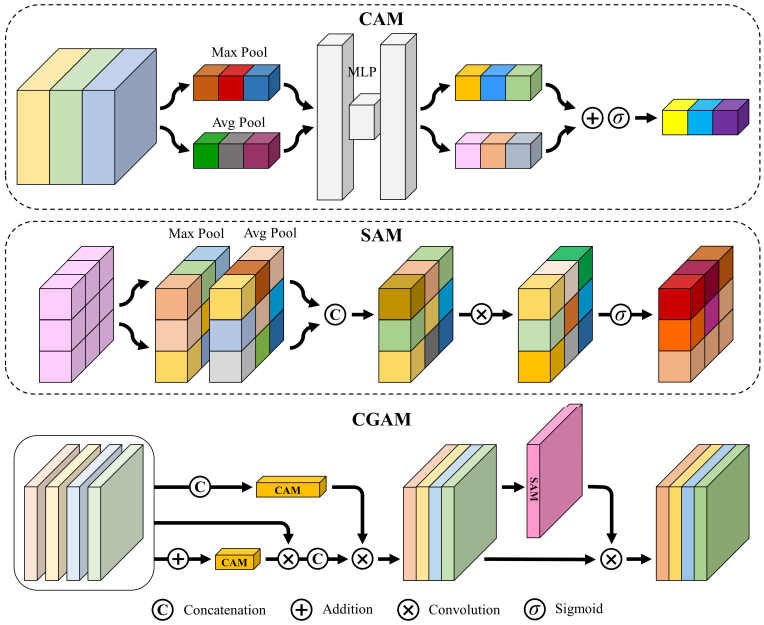

Figure 5. The architecture of CGAM.

Through the CGAM, refined feature maps with spatial attention are generated. As a complement to CGAM, the spatial attention module (SAM) [38] is added to focus on the more precise position of the semantic and location information on the feature map. Finally, a feature map with channel refinement and spatial refinement features is obtained, and the final change map $C_{map} \in \mathbb{R}^{H \times W \times 2}$ is generated after passing it through a 1×1 convolutional layer.

3.4. Multi-SUAV Collaboration Platform

A multi-UAV collaboration platform is established for change detection of nature reserve garbage scattered areas. The detection and localization tasks can be run on a computer to reduce the requirements for SUAVs' energy consumption and performance. At the same time, thanks to its openness, the platform can simultaneously connect multiple SUAVs for data transmission. The platform can realize the path planning of multiple SUAVs on a visual map interface, and display the video stream as well as flight information returned by multiple SUAVs in real time. Meanwhile, the location information of the detected garbage scattered areas can be read and then marked on the map. The visual interface of the platform is shown in Figure 6. In function, the platform is mainly composed of a multi-SUAV collaboration module, path planning module and location module.

3.4.1. Multi-SUAV Collaboration

The multi-SUAV collaboration module is used to build a multi-SUAV collaboration remote sensing control system with open interfaces. The system can connect multiple SUAVs at the same time, receive the video streams returned by each SUAV and display them in real-time. Meanwhile, the cruise path of each SUAV can be planned by manual punctuation in the visual map interface, and the real-time flight record of each SUAV can also be displayed in the form of a text stream.

The DJI's civilian-grade SUAVs are chosen as the video acquisition terminal of our multi-SUAV platform because DJI provides a stable software development kit (SDK) called DJI Mobile SDK, which is convenient for our system development according to practical applications [46–48]. Through this SDK, the flight data of the UAVs can be accessed in real-time to realize functions such as automatic cruise, gimbal remote control, real-time video streaming transmission, real-time GPS information acquisition and SUAVs' status monitoring.

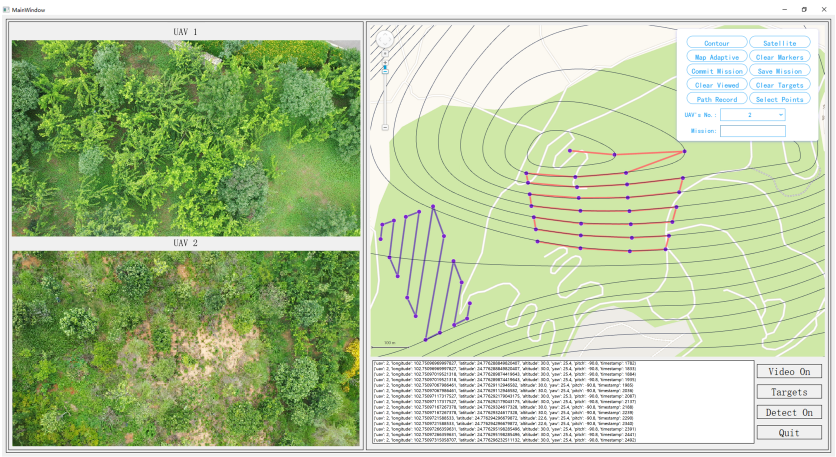

Figure 6. The multi-UAV collaboration platform.

3.4.2. Path Planning

The path planning module is designed for the global planning of flight range and route. Path planning is essentially a waypoint task, which is needed to point the waypoint path on the map first, and then achieved through UAVs passing through the ordered GPS coordinates with elevation information in turn [49,50]. The path planning method, mainly based on satellite maps and supplemented by the digital elevation model (DEM) [51–53], is adopted, which not only realizes safer UAV route planning in nature reserves but also makes the location of garbage-scattered areas more accurate.

3.4.3. Location

The location module is mainly based on GPS information and coordinate transformation, which is the core of the garbage-scattered area location. Using the GPS information [54,55] contained in the video stream transmitted back by UAVs, the pixel coordinates of the detected change area in each video frame are converted into GPS coordinates so as to realize the location of the garbage-scattered areas.

3.5. Training

3.5.1. Loss Function

Our network, which combines the image registration and change detection task, is trained end-to-end. The pre-trained ResNet18 feature extractor backbone is unfrozen and participates in the training to update parameters. According to the different tasks performed at each stage, the loss function of the proposed network is divided into the loss function of the image registration phase and the loss function of the change detection phase.

In the field of image registration, the endpoint error (EPE), which is the standard error measurement for optical flow estimation, is used as the training loss. It is the displacement between the estimated optical flow and the ground truth, calculated from the Euclidean distance. As proposed in FlowNet [25], we utilize the optical flow field information of different optical flow pyramid layers for multi-scale training loss. The multi-scale EPE loss can be formulated as:

$$\mathcal{L}_{MEPE} = \sum_{l=1}^{L} \alpha^l \frac{1}{H^l \times W^l} \sum_{\mathbf{c}}^{H^l \times W^l} \|\mathbf{w}_{est}^l(\mathbf{c}) - \mathbf{w}_{gt}^l(\mathbf{c})\|_2 \tag{14}$$

where l denotes the level of the L-level optical flow pyramid. H^l and W^l are the feature map size in the level l. $\mathbf{w}_{est}^l$ and $\mathbf{w}_{gt}^l$ respectively denote the optical flow field estimated by the network at the l pyramid level and the corresponding ground truth field. $\mathbf{c} \in \mathbb{Z}^2$ is the coordinate in the optical flow field. α^l is the weight coefficient of each pyramid level to adjust the weight of different pyramid layers.

In the phase of the image registration, the focal loss [56] is utilized as the loss function to solve the sample imbalance problem in that the number of the changed pixel is much less than the number of the unchanged pixel. In addition, the dice coefficient is also added to assist in tackling this problem. Formally, the combination loss function of change detection is defined as follows:

$$\mathcal{L}_{CD} = \mathcal{L}_{focal} + \mathcal{L}_{dice} \tag{15}$$

$$\mathcal{L}_{focal} = \frac{1}{H \times W} \sum_{k=1}^{H \times W} \Big(-(1 - p_t)^{\gamma} log(p_t) \Big) \tag{16}$$

$$\mathcal{L}_{dice} = \frac{1}{H \times W} \sum_{k=1}^{H \times W} \left(1 - \frac{2 \cdot Y \cdot softmax(\hat{Y})}{Y + softmax(\hat{Y})} \right) \tag{17}$$

where p_t is the simplified form of the cross entropy loss, and γ is the modulating factor. Y denotes the ground truth and $\hat{Y}$ denotes the predicted change map.

The final loss function $\mathcal{L}$ is the addition of $\mathcal{L}_{MEPE}$ and $\mathcal{L}_{CD}$, which is defined as:

$$\mathcal{L} = \mathcal{L}_{CD} + \beta \cdot \mathcal{L}_{MEPE} \tag{18}$$

where β is a hyperparameter. The final task of the entire network is to perform change detection, however, the accuracy of this task is affected by the effect of the previous registration task. Therefore, the hyperparameter β is added to adjust the size of the loss function $\mathcal{L}_{MEPE}$, so as to weigh the proportion between the two tasks, which in turn optimizes the final change detection effect.

3.5.2. Dataset Generation

Our network requires the supervised training data consisting of image pairs, optical flow pairs, warp optical flow and ground truth of change detection. Unlike other vision tasks, the used dataset of low-altitude remote sensing garbage in nature reserves contains not only the ground truth of warp optical flow for image registration but also the ground truth of garbage change detection. In general, obtaining such a dataset is hard, and no public datasets exist that can satisfy the proposed network's requirements for the dataset. Therefore, we decide to make a dataset for training and validation of the network.

As shown in Figure 7, the artificial dataset is generated according to the following steps: (1) The high-resolution images of the nature reserve ground are shot vertically downward by SUAVs at a height of 30 m. (2) A large number of common garbage images are collected, and their backgrounds are removed through matting software to generate massive garbage image patches. (3) Some garbage patches are selected randomly to form several garbage distribution areas with a certain probability distribution, and these areas are covered randomly into the original image to generate the source image with a small amount of garbage scattered areas, and the corresponding binary change map. (4) After the affine transformation on the original image, the target image is generated by covering more garbage scattered areas on it, and the binary change map is also produced. At the same time, the affine matrix is saved and the optical flow map is generated through this matrix to record the perspective change from the source image to the target image. (5) The change map of the target image is restored to the perspective of the source image through

the inverse matrix of the affine matrix and added with the change map of the source image to generate the change map ground truth of the source and target images. (6) At the same position on a group of source, target, change and optical flow maps, a set of dataset images is obtained through cropping with small square boxes of the same size. Multiple sets of dataset images can be obtained by performing the above operations at random positions multiple times on a group of source images. (7) The entire dataset is obtained by deleting those image sets with too large black border areas and repeated cropping areas, after performing the above step on all source image groups. (8) To consummate the dataset, data augmentation is used for the target images to increase the difference from the source images and enhance the robustness of the network after training.

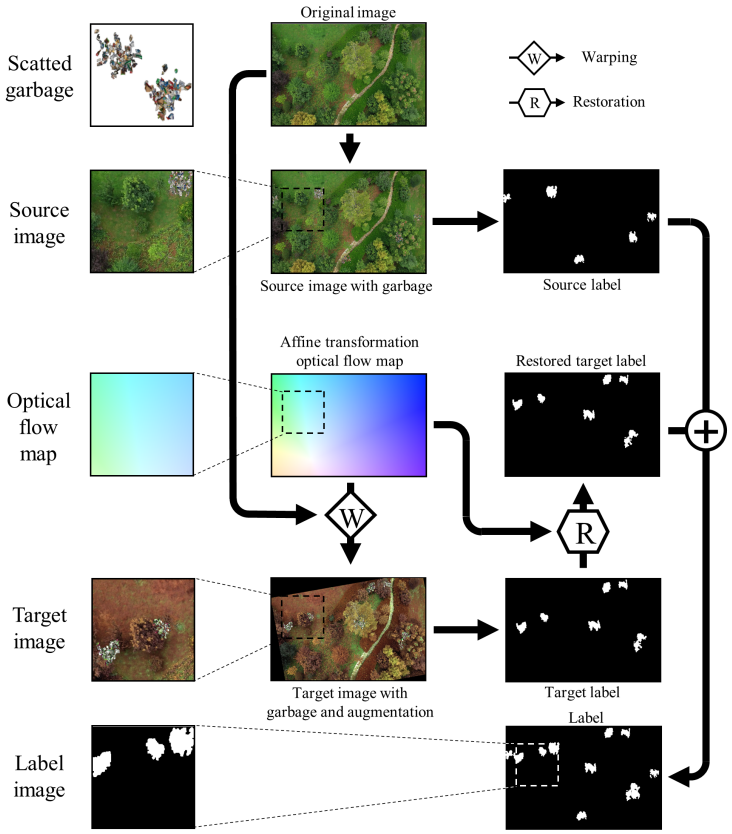

Figure 7. The procedure of dataset generation.

4. Experiments

4.1. Training Datasets

To evaluate our method, following the steps introduced in Section 3.5.2, we generated a bi-temporal SUAV low-altitude remote sensing dataset for garbage-scattered areas change detection in nature reserves, and designed a series of experiments on this dataset. To increase the diversity of training data, we use several data augmentation methods when generating the dataset. The data augmentation applied to the original images in step 4 is mainly affine transformation, including rotation, translation and scaling. In step 8, the data augmentation applied in target images is Gaussian blur, Gaussian noise and color jitter. The parameter settings of data augmentation are shown in Table 1. We used SUAVs to capture 2000 original ground images with a resolution of 3840 × 2160 pixels in nature reserves and obtained a dataset containing 48,000 pairs of images with a resolution of 256 × 256 pixels through the above operations. Then, the dataset is divided into a training set, validation set and testing set in a ratio of 6:2:2. Some selected samples from the dataset are shown in Figure 8.

Table 1. Parameter setting of data augmentation.

Parameter Name	Value
rotation range	±30°
translation range	±0.2
scale range	0.8–1.2
pixel intensity range	0.5–1.5 per channel
contrast range	0.5–2 per channel
Gaussian distribution	$\mu = 0, \sigma = 0.05 \times 255$
Gaussian blur probability	0.5
Gaussian noise probability	0.5 per channel
Gaussian kernel size	3

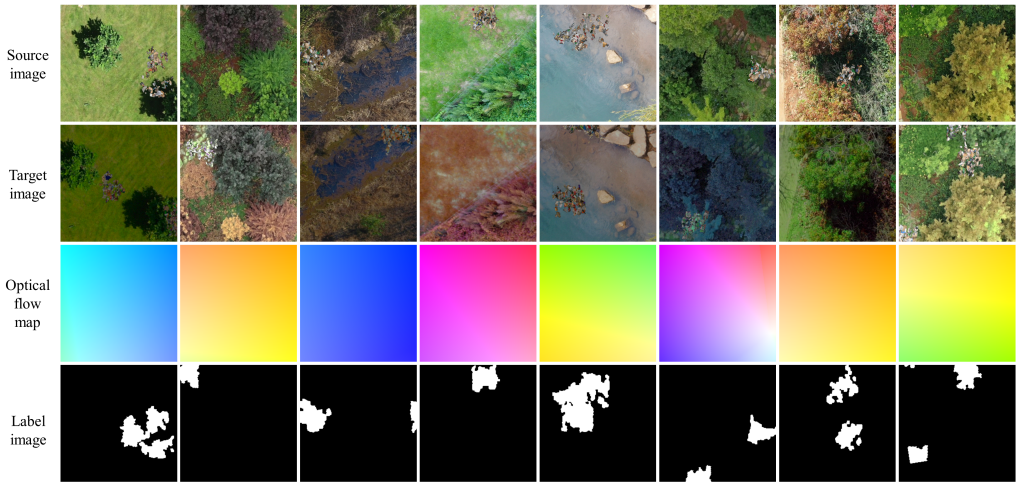

Figure 8. The overview of the dataset. Optical flow field color is encoded with the same method used in FlowNet [25].

4.2. Evaluation Metrics

To evaluate the performance of the proposed method, we used three widely used evaluation criteria: Precision (P), Recall (R), and F_1 Score (F_1). The formula of P, R and F_1 are defined as follows:

$$P = \frac{T_P}{T_P + F_P} \tag{19}$$

$$R = \frac{T_P}{T_P + F_N} \tag{20}$$

$$F_1 = \frac{2PR}{P + R} \tag{21}$$

where F_P (False Positive) and F_N (False Negative) represent the number of incorrectly detected changes, and T_P (True Positive) represents the number of correctly detected ones. $T_P + F_P$ is the total number of detected changes, and $T_P + F_N$ is the total number of actual changes. F_1 is the harmonic mean of P and R. In the change detection task, the higher precision value denotes fewer false detection results, and the higher recall value denotes fewer missed detection results.

4.3. Implementation Details

All of our experiments were based on the deep learning framework Pytorch and conducted on a single NVIDIA RTX-3090. To train our method, the Adam optimizer [57] was adopted with an initial learning rate of 1×10^{-4} for optimization. The momentum is set to 4×10^{-4} and the weight decay is set to 4×10^{-4}. The learning rate decays by 0.5 times every 10 epochs starting from the 60th epoch. The Kaiming normalization is used

to initialize the weights of each convolutional layer. The hyperparameter β was set to 1×10^{-3}. The weights coefficients α are set to $\alpha_1 = 0.005$, $\alpha_2 = 0.01$, $\alpha_3 = 0.02$, $\alpha_4 = 0.08$, $\alpha_5 = 0.32$. Except for our method, all other models are trained with the default parameters of their official code. Model training is finished at 100 epochs with a batch size of 16.

5. Results

To evaluate the performance of the proposed method, we make a comparison with the several representative and SOTA change detection methods on our dataset.

- FC-Siam-Conc [58]: The baseline model for change detection, which is fully consisted of convolution. It is a simple combination of UNet and Siamese networks and uses feature concatenation to fuse the bi-temporal information.
- FC-Siam-Diff [58]: The baseline model for change detection, whose architecture is similar to FC-Siam-Conc, but uses multi-scale feature difference to fuse the bi-temporal information.
- UNet++_MSOF [59]: Feature fusion method, which inputs concatenated bi-temporal images into UNet++, and uses the multiple side output fusion for deep supervision.
- IFN [37]: Multi-scale feature concatenation method, which fuses the multi-level deep features of images with different features by attention modules, and uses a deep supervision strategy for optimization.
- DASNet [12]: Attention-based method, which extracts features by a Siamese backbone, and uses a dual attention mechanism to build connections between local features to obtain more discriminant feature representations.
- BIT [60]: Transformer-based method, which models contexts within the spatial-temporal domain through multi-attention heads, and projects them to the pixel space to refine the representation of the original features.
- SNUNet-CD [40]: Multi-scale feature concatenation method, which combines UNet++ and Siamese network, and uses the ensemble channel attention module to integrate multi-level outputs to perform deep supervision.
- RDP-Net [61]: Feature fusion method, which uses region detail preserving the network to improve the detection performance on boundaries and small regions.

We implement the above CD methods using their public codes with default hyperparameters. We train the above networks on our dataset to examine their change detection performance in realistic bi-temporal SUAV low-altitude remote sensing images with viewpoint changes. Furthermore, as a comparison, these CD networks are also trained on our dataset with registered bi-temporal images. The bi-temporal images are pre-registered by the outstanding optical flow registration network GLU-Net [62], and we use the optical flow in our dataset for the GLU-Net training.

Table 2 reports the overall comparisons of detection accuracy, parameters number and FPS on our dataset. Our proposed RegCD-Net can outperform the other change detection methods with relatively few parameters, fast speed and better integration without additional registration networks. In comparison with UNet++_MSOF, IFN, BIT and SNUNet-CD, our method achieves the highest P (96.74%), R (95.92%) and F1 (96.32%) with minimal parameters (20.66 M). Owing to simple network structure, FC-Siam-conc, FC-Siam-diff and RDP-Net have the least parameters, and the former two, also have the fastest FPS. By contrast, our RegCD-Net achieves at least 1% accuracy improvement with only about 5M more parameters and 0.7 less FPS. In addition, although DASNet gets the highest R (96.42%), our RegCD-Net achieves the second best R (slightly lower 0.5%) only with 32% parameters of DASNet.

Table 2. Performance comparison results on our dataset. The parameters of different methods with pre-registration contain the parameters of GLU-Net (13.59 M). The FPS is computed by adding the registration speed of GLU-Net and the detection speed of different methods.

Method	Pre-Registration	Backbone	P (%)	R (%)	F1 (%)	Params (M)	FPS
FC-Siam-conc	×	UNet	91.97	57.06	70.24	1.55	×
FC-Siam-diff	×	UNet	92.20	56.69	70.02	**1.35**	×
UNet++_MSOF	×	UNet++	85.90	79.76	82.53	8.83	×
IFN	×	VGG16	94.88	57.75	71.8	35.99	×
DASNet	×	ResNet50	45.79	79.37	64.29	50.27	×
BIT	×	ResNet18	88.18	76.63	81.43	12.40	×
SNUNet-CD	×	UNet++	87.59	67.07	75.28	12.03	×
RDP-Net	×	RSNet	87.92	85.62	86.73	1.70	×
FC-Siam-conc	✓	UNet	95.94	94.31	94.67	15.14 (1.55 + 13.59)	**26.3**
FC-Siam-diff	✓	UNet	94.78	92.71	93.73	**14.94** (1.35 + 13.59)	26.0
UNet++_MSOF	✓	UNet++	95.28	94.16	94.72	22.42 (8.83 + 13.59)	25.1
IFN	✓	VGG16	95.36	94.81	95.08	49.58 (35.99 + 13.59)	21.2
DASNet	✓	ResNet50	85.42	**96.42**	90.59	63.86 (50.27 + 13.59)	18.9
BIT	✓	ResNet18	93.35	95.29	94.31	25.99 (12.40 + 13.59)	20.1
SNUNet-CD	✓	UNet++	96.40	95.12	95.74	25.62 (12.03 + 13.59)	20.7
RDP-Net	✓	RSNet	95.23	94.80	95.01	15.29 (1.70 + 13.59)	22.0
RegCD-Net (our)	×	ResNet18	**96.74**	95.92	**96.32**	20.66	25.6

The visualization comparison of different methods on our dataset is displayed in Figure 9. The true positive, true negative, false positive and false negative are indicated by white, black, red and green, respectively in the figure. From Figure 9, we can observe that our RegCD-Net achieves better detection performance than others, which mainly benefited from three perspectives. Firstly, our RegCD-Net employs global and local correlations in the deepest and shallow layers to generate optical flow, respectively, to achieve better registration performance in bi-temporal images with large viewpoint changes, thereby indirectly improving change detection accuracy. Secondly, our RegCD-Net utilizes nested connections in the up-sampling process, which combines rich semantic information and precise location information, to achieve more delicate edge performance in the change map. Furthermore, we also use the attention module CGAM to fuse features of the different semantic level paths to automatically emphasize more precise change edge representation. Thirdly, our RegCD-Net integrates registration and change detection sub-networks into a single network, enabling end-to-end optimization of change detection on bi-temporal low-altitude remote sensing images with viewpoint changes, which can reduce the influence of imprecise pre-registration results on change detection performance. Benefiting from an end-to-end structure, our method also has good real-time performance.

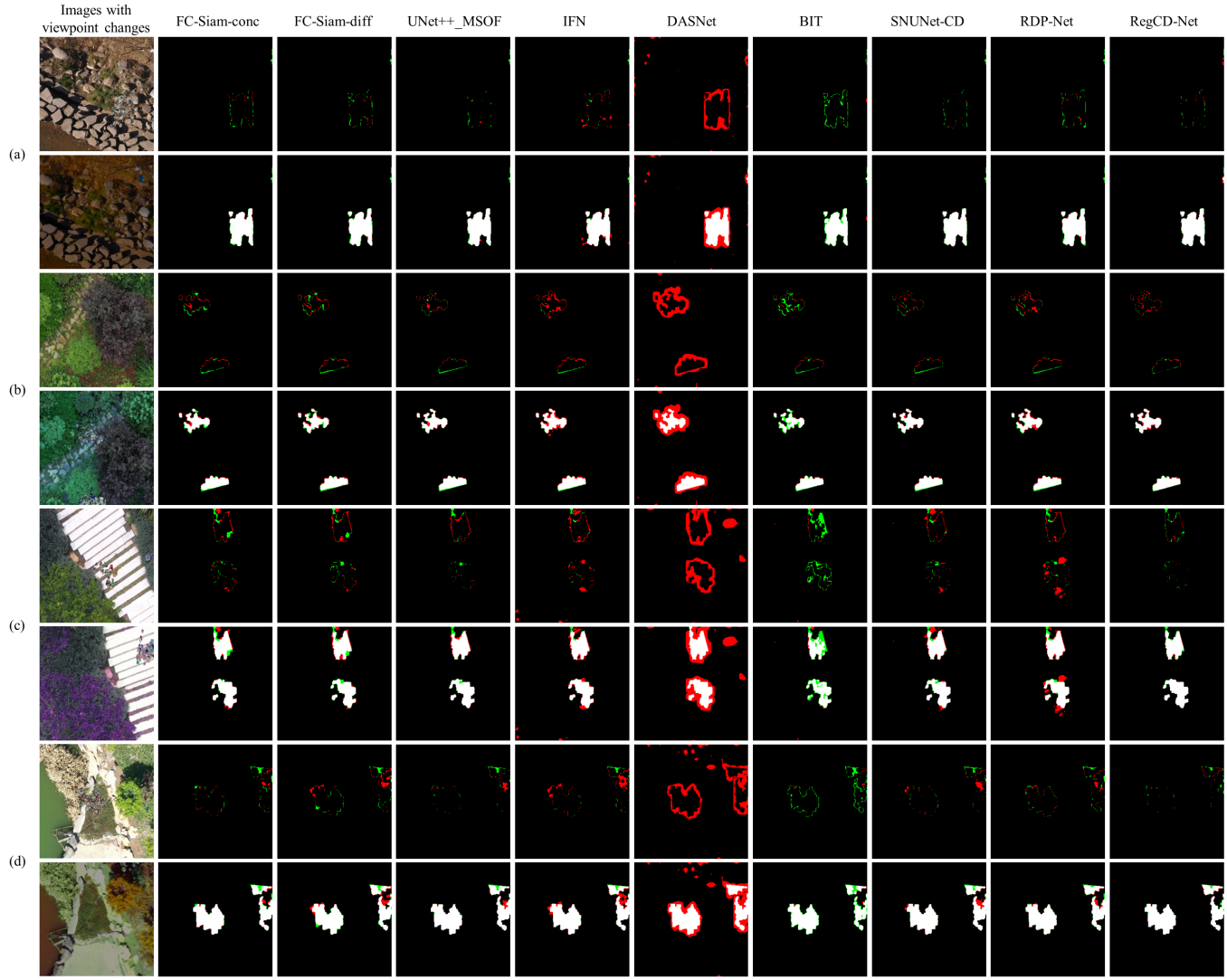

Figure 9. Visualization results of different methods on our dataset. Four groups of different bi-temporal images are marked with (**a**–**d**). Different colors are used for a better view, which is white for true positive, black for true negative, red for false positive and green for false negative. The change maps in the first row of each group only contain false positives and false negatives for better edge error representation.

6. Ablation Studies

6.1. Ablation on Loss Function

We perform ablation on the loss function to validate its performance in our end-to-end network RegCD-Net, where hyperparameter β is the core parameter in connecting two task losses of optical flow registration and change detection. Table 3 shows the ablation study of the hyperparameter β on our made dataset. From Table 3, we observe consistent drops in AEPE as the decreasing of the β, while the peaks of P, R and F1 emerge at $\beta = 1 \times 10^{-3}$. That is because when adjusting the β of the two loss functions, more weight on registration loss $\mathcal{L}_{MEPE}$ means better loss optimization on it and more accurate registration performance. Better registration performance further improves the accuracy of the change detection. However, more weight on registration loss also means that network is more inclined to the registration task. That is to say, when the loss decreases, it will be insensitive to the change detection loss changes, which makes it difficult to effectively optimize the change detection task, resulting in decreasing after accuracy reaches the peak value. We choose the value of β at the accuracy peak as the loss function hyperparameter for subsequent

RegCD-Net ablation experiments. The results of optical flow registration at chosen β are shown in Figure 10.

Table 3. Ablation study of hyperparameter β on our dataset.

β	AEPE	P (%)	R (%)	F1 (%)
1×10^{-2}	**4.20**	95.42	95.35	95.38
1×10^{-3}	5.22	**96.74**	**95.92**	**96.32**
1×10^{-4}	6.15	95.28	94.07	94.67
1×10^{-5}	7.34	94.57	92.35	93.45
1×10^{-6}	8.48	94.15	90.66	92.37
1×10^{-7}	9.55	93.68	88.87	91.14

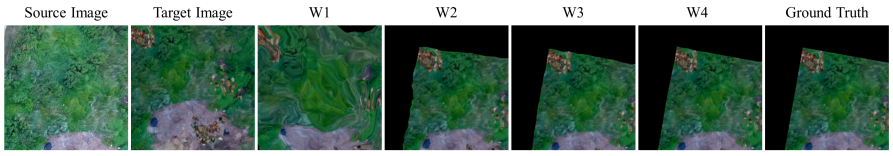

Figure 10. The registration results in $\beta = 1 \times 10^{-3}$. W_i denotes different optical flow pyramid levels.

6.2. Effect of Assemble Correlation and Nested Connection

We perform ablation on the optical flow correlation layer and up-sampling connection module. We complete our experiments by adding or deleting the global correlation layer and replacing connection methods. All models in the experiments do not contain any attention modules. As shown in Table 4, only local correlation and UNet connection perform the worst performance. When replacing the UNet connection with nested connections, the accuracy has an improvement, which indicates that nested connections contribute to change detection performance improvement. We also observe that compared to the model with only local correlation and Unet connection, the model additionally containing global correlation has obvious accuracy improvement, especially in recall score. It may be because global correlation performs better in large displacement optical flow registration, and better registration results will directly lead to higher detection accuracy. Compared to the above, our RegCD-Net, which contains two kinds of correlation layer and nested connection can perceive large-scale displacement under the premise of ensuring local registration effect and achieve fusing shallow location information and deep semantic information, which will undoubtedly perform the best in P, R and F1 score.

Table 4. Ablation study of our RegCD-Net on our dataset. Ablations are local correlation (LC), global correlation (GC), Unet connection (UC) and nested connection (NC). All of the following models do not add attention.

LC	GC	UC	NC	P (%)	R (%)	F1 (%)
✓	×	✓	×	91.33	86.55	88.82
✓	×	×	✓	93.24	88.98	91.00
✓	✓	✓	×	93.81	91.50	92.63
✓	✓	×	✓	**94.88**	**94.02**	**94.45**

6.3. Comparison on Attention Modules

To further evaluate the performance of our RegCD-Net, we add the attention mechanism after four up-sampling outputs for performing network deep supervision, to automatically select better feature representations from different semantic level up-sampling pathways. We designed ablation experiments on attention modules to compare the performance of the widely used attention module CBAM and our designed CGAM on our dataset. Table 5 shows that the model, which only uses CAM or SAM, has small improvements over the no attention module, and the model combining two of the above performs better

in accuracy. By contrast, the model adding our CGAM further improves the accuracy. It proves that our CGAM can effectively focus on the intra-group common feature of four up-sampling outputs, and selects more appropriate feature representation channels between four groups to make the boundary position in the change detection results more precise.

Table 5. Ablation study of attention models on RegCD-Net with our dataset. Ablations are the channel attention module (CAM), spatial attention module (SAM) and channel group attention module (CGAM).

CAM	SAM	CGAM (Our)	P (%)	R (%)	F1 (%)
×	×	×	94.88	94.02	94.45
✓	×	×	95.26	94.74	94.99
×	✓	×	94.96	94.93	94.94
✓	✓	×	95.82	94.96	95.38
×	×	✓	**96.74**	**95.92**	**96.32**

7. Discussion

The above extensive comparison experiments and ablation studies demonstrate the effectiveness of proposed RegCD-Net. Specifically, several observations from the above experiments are summarized as follows:

- RegCD-Net can achieve the registration of bi-temporal SUAV low-altitude remote sensing images with viewpoint changes, with no pre-registration, better integrity, fast speed and relatively fewer parameter numbers. It can improve change detection accuracy through end-to-end optimization.
- Global and local assemble correlation can capture large displacements between feature maps in the deep layers of the network, and can achieve more detailed local associations in the shallow layers. It achieves fine registration of remote sensing images with large viewpoint changes by the construction of a coarse-to-fine optical flow pyramid.
- Nested connection up-sampling can combine the rich semantic information in the deep layers and the precise location information in the shallow layers through different skip connections and up-sampling pathways to achieve more delicate edge detection performance in the change detection maps.
- Channel group attention module CGAM can effectively focus on the intra-group common feature and the inter-group different features of four up-sampling outputs from different up-sampling pathways, and select more appropriate feature representation channels to make the boundary position in the change detection results more precise.

These observations can provide us with very meaningful guidance, laying the foundation for our subsequent construction of faster, fewer parameters and better performance in end-to-end change detection network in the field of low-altitude SUAV remote sensing with a higher-resolution and wider view.

8. Conclusions

In this paper, we propose RegCD-Net, an end-to-end method for change detection in bi-temporal SUAV low-altitude remote sensing images. We integrate registration and change detection sub-networks in a network, and utilize assemble correlation, nested connection and deep supervision attention module to optimize the uncertainty of pixels at the edge of change detection results. In addition, we also generate a bi-temporal SUAV low-altitude remote sensing dataset for change detection of garbage-scattered areas in nature reserves and experiment on this dataset. Extensive experiments have validated the effectiveness of our method, without an additional pre-registration network. Our end-to-end RegCD-Net can detect more precise edge changes with the best accuracy, fewer network parameters and better detection speed. In the future, we will focus on high-level simplification and integration of network structures in end-to-end no pre-registration change detection, and

apply our method to bi-temporal low-altitude remote sensing images with higher resolution and larger view field.

Author Contributions: K.Y.: writing—original draft preparation, writing—review and editing, conceptualization, methodology, software, validation, data curation, visualization. Y.D.: methodology, software, data curation, visualization. Y.Y.: writing—review and editing, investigation, funding acquisition. L.X.: investigation, supervision. 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 (41971392) and Yunnan Province Ten-thousand Talents Program.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The author would like to thank the experimental platform provided by the Laboratory of Pattern Recognition and Artificial Intelligence and also thank the public code provided by the authors of different change detection methods on GitHub.

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

References

1. Xiao, P.; Zhang, X.; Wang, D.; Yuan, M.; Feng, X.; Kelly, M. Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 402–414. [CrossRef]
2. Gao, S.; Li, W.; Sun, K.; Wei, J.; Chen, Y.; Wang, X. Built-Up Area Change Detection Using Multi-Task Network with Object-Level Refinement. *Remote Sens.* **2022**, *14*, 957. [CrossRef]
3. Xing, J.; Sieber, R.; Caelli, T. A scale-invariant change detection method for land use/cover change research. *ISPRS J. Photogramm. Remote Sens.* **2018**, *141*, 252–264. [CrossRef]
4. Lv, Z.; Liu, T.; Zhang, P.; Atli Benediktsson, J.; Chen, Y. Land cover change detection based on adaptive contextual information using bi-temporal remote sensing images. *Remote Sens.* **2018**, *10*, 901. [CrossRef]
5. Lu, M.; Pebesma, E.; Sanchez, A.; Verbesselt, J. Spatio-temporal change detection from multidimensional arrays: Detecting deforestation from MODIS time series. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 227–236. [CrossRef]
6. Vega, P.J.S.; da Costa, G.A.O.P.; Feitosa, R.Q.; Adarme, M.X.O.; de Almeida, C.A.; Heipke, C.; Rottensteiner, F. An unsupervised domain adaptation approach for change detection and its application to deforestation mapping in tropical biomes. *ISPRS J. Photogramm. Remote Sens.* **2021**, *181*, 113–128. [CrossRef]
7. Jiang, J.; Xing, Y.; Wei, W.; Yan, E.; Xiang, J.; Mo, D. DSNUNet: An Improved Forest Change Detection Network by Combining Sentinel-1 and Sentinel-2 Images. *Remote Sens.* **2022**, *14*, 5046. [CrossRef]
8. Niu, C.; Zhang, H.; Liu, W.; Li, R.; Hu, T. Using a fully polarimetric SAR to detect landslide in complex surroundings: Case study of 2015 Shenzhen landslide. *ISPRS J. Photogramm. Remote Sens.* **2021**, *174*, 56–67. [CrossRef]
9. Wang, X.; Fan, X.; Xu, Q.; Du, P. Change detection-based co-seismic landslide mapping through extended morphological profiles and ensemble strategy. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 225–239. [CrossRef]
10. Tang, L.; Deng, Y.; Ma, Y.; Huang, J.; Ma, J. SuperFusion: A Versatile Image Registration and Fusion Network with Semantic Awareness. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 2121–2137. [CrossRef]
11. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
12. Chen, J.; Yuan, Z.; Peng, J.; Chen, L.; Huang, H.; Zhu, J.; Liu, Y.; Li, H. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1194–1206. [CrossRef]
13. Chen, P.; Guo, L.; Zhang, X.; Qin, K.; Ma, W.; Jiao, L. Attention-Guided Siamese Fusion Network for Change Detection of Remote Sensing Images. *Remote Sens.* **2021**, *13*, 4597. [CrossRef]
14. Zhang, L.; Hu, X.; Zhang, M.; Shu, Z.; Zhou, H. Object-level change detection with a dual correlation attention-guided detector. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 147–160. [CrossRef]
15. Cheng, H.; Wu, H.; Zheng, J.; Qi, K.; Liu, W. A hierarchical self-attention augmented Laplacian pyramid expanding network for change detection in high-resolution remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *182*, 52–66. [CrossRef]
16. Shen, Q.; Huang, J.; Wang, M.; Tao, S.; Yang, R.; Zhang, X. Semantic feature-constrained multitask siamese network for building change detection in high-spatial-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *189*, 78–94. [CrossRef]
17. Chen, P.; Zhang, B.; Hong, D.; Chen, Z.; Yang, X.; Li, B. FCCDN: Feature constraint network for VHR image change detection. *ISPRS J. Photogramm. Remote Sens.* **2022**, *187*, 101–119. [CrossRef]

18. Zhu, Q.; Guo, X.; Deng, W.; Guan, Q.; Zhong, Y.; Zhang, L.; Li, D. Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 63–78. [CrossRef]
19. Zhang, X.; He, L.; Qin, K.; Dang, Q.; Si, H.; Tang, X.; Jiao, L. SMD-Net: Siamese Multi-Scale Difference-Enhancement Network for Change Detection in Remote Sensing. *Remote Sens.* **2022**, *14*, 1580. [CrossRef]
20. 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]
21. Li, J.; Zhu, S.; Gao, Y.; Zhang, G.; Xu, Y. Change Detection for High-Resolution Remote Sensing Images Based on a Multi-Scale Attention Siamese Network. *Remote Sens.* **2022**, *14*, 3464. [CrossRef]
22. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [CrossRef]
23. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2017–2025.
24. 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; pp. 234–241.
25. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
26. Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; Brox, T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2462–2470.
27. Ranjan, A.; Black, M.J. Optical flow estimation using a spatial pyramid network. In Proceedings of the IEEE conference on computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4161–4170.
28. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8934–8943.
29. Melekhov, I.; Tiulpin, A.; Sattler, T.; Pollefeys, M.; Rahtu, E.; Kannala, J. Dgc-net: Dense geometric correspondence network. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 1034–1042.
30. Rocco, I.; Arandjelovic, R.; Sivic, J. Convolutional neural network architecture for geometric matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6148–6157.
31. Xiao, X.; Lian, S.; Luo, Z.; Li, S. Weighted res-unet for high-quality retina vessel segmentation. In Proceedings of the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 19–21 October 2018; pp. 327–331.
32. Guan, S.; Khan, A.A.; Sikdar, S.; Chitnis, P.V. Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal. *IEEE J. Biomed. Health Inform.* **2019**, *24*, 568–576. [CrossRef] [PubMed]
33. 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, 26 June–1 July 2016; pp. 770–778.
34. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
35. Zheng, Z.; Wan, Y.; Zhang, Y.; Xiang, S.; Peng, D.; Zhang, B. CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 247–267. [CrossRef]
36. Bi, H.B.; Lu, D.; Zhu, H.H.; Yang, L.N.; Guan, H.P. STA-Net: Spatial-temporal attention network for video salient object detection. *Appl. Intell.* **2021**, *51*, 3450–3459. [CrossRef]
37. Zhang, C.; Yue, P.; Tapete, D.; Jiang, L.; Shangguan, B.; Huang, L.; Liu, G. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *166*, 183–200. [CrossRef]
38. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
39. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–11.
40. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
41. Hui, T.W.; Tang, X.; Loy, C.C. A lightweight optical flow CNN—Revisiting data fidelity and regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 2555–2569. [CrossRef]
42. Hui, T.W.; Tang, X.; Loy, C.C. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8981–8989.
43. Kim, S.; Min, D.; Jeong, S.; Kim, S.; Jeon, S.; Sohn, K. Semantic attribute matching networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12339–12348.

44. Rocco, I.; Cimpoi, M.; Arandjelović, R.; Torii, A.; Pajdla, T.; Sivic, J. Neighbourhood consensus networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
45. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
46. Xing, L.; Fan, X.; Dong, Y.; Xiong, Z.; Xing, L.; Yang, Y.; Bai, H.; Zhou, C. Multi-UAV cooperative system for search and rescue based on YOLOv5. *Int. J. Disaster Risk Reduct.* **2022**, *76*, 102972. [CrossRef]
47. Stanković, M.; Mirza, M.M.; Karabiyik, U. UAV forensics: DJI mini 2 case study. *Drones* **2021**, *5*, 49. [CrossRef]
48. Salamh, F.E.; Mirza, M.M.; Karabiyik, U. UAV forensic analysis and software tools assessment: DJI Phantom 4 and Matrice 210 as case studies. *Electronics* **2021**, *10*, 733. [CrossRef]
49. Kwak, J.; Sung, Y. Autonomous UAV flight control for GPS-based navigation. *IEEE Access* **2018**, *6*, 37947–37955. [CrossRef]
50. Grayson, B.; Penna, N.T.; Mills, J.P.; Grant, D.S. GPS precise point positioning for UAV photogrammetry. *Photogramm. Rec.* **2018**, *33*, 427–447. [CrossRef]
51. Annis, A.; Nardi, F.; Petroselli, A.; Apollonio, C.; Arcangeletti, E.; Tauro, F.; Belli, C.; Bianconi, R.; Grimaldi, S. UAV-DEMs for small-scale flood hazard mapping. *Water* **2020**, *12*, 1717. [CrossRef]
52. Ajayi, O.G.; Salubi, A.A.; Angbas, A.F.; Odigure, M.G. Generation of accurate digital elevation models from UAV acquired low percentage overlapping images. *Int. J. Remote Sens.* **2017**, *38*, 3113–3134. [CrossRef]
53. Uysal, M.; Toprak, A.S.; Polat, N. DEM generation with UAV Photogrammetry and accuracy analysis in Sahitler hill. *Measurement* **2015**, *73*, 539–543. [CrossRef]
54. Xi, X.; Cao, X.; Yang, P.; Chen, J.; Quek, T.; Wu, D. Joint user association and UAV location optimization for UAV-aided communications. *IEEE Wirel. Commun. Lett.* **2019**, *8*, 1688–1691. [CrossRef]
55. Duffy, J.P.; Cunliffe, A.M.; DeBell, L.; Sandbrook, C.; Wich, S.A.; Shutler, J.D.; Myers-Smith, I.H.; Varela, M.R.; Anderson, K. Location, location, location: Considerations when using lightweight drones in challenging environments. *Remote Sens. Ecol. Conserv.* **2018**, *4*, 7–19. [CrossRef]
56. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 24–27 October 2017; pp. 2980–2988.
57. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
58. Daudt, R.C.; Le Saux, B.; Boulch, A. Fully convolutional siamese networks for change detection. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4063–4067.
59. Peng, D.; Zhang, Y.; Guan, H. End-to-end change detection for high resolution satellite images using improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
60. Chen, H.; Qi, Z.; Shi, Z. Remote sensing image change detection with transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–14. [CrossRef]
61. Chen, H.; Pu, F.; Yang, R.; Tang, R.; Xu, X. RDP-Net: Region Detail Preserving Network for Change Detection. *arXiv* **2022**, arXiv:2202.09745.
62. Truong, P.; Danelljan, M.; Timofte, R. GLU-Net: Global-local universal network for dense flow and correspondences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6258–6268.