remote sensing

MDPI

Article

Self-Supervised Stereo Matching Method Based on SRWP and PCAM for Urban Satellite Images

Wen Chen, Hao Chen * and Shuting Yang

School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150006, China; 19b905005@stu.hit.edu.cn (W.C.); 21b905002@stu.hit.edu.cn (S.Y.)
* Correspondence: hit_hao@hit.edu.cn

Abstract: In this paper, we propose a self-supervised stereo matching method based on superpixel random walk pre-matching (SRWP) and parallax-channel attention mechanism (PCAM). Our method is divided into two stages, training and testing. First, in the training stage, we obtain pre-matching results of stereo images based on superpixel random walk, and some matching points with high confidence are selected as labeled samples. Then, a stereo matching network is constructed to describe the matching correlation by calculating the attention scores of any two points between different images through the parallax-channel attention mechanism, superimposing the scores of each layer to calculate the disparity. The network is trained using the labeled samples and some unsupervised constraint criteria. Finally, in the testing stage, the trained network is used to obtain stereo matching relations of stereo images. The proposed method does not need manually labeled training samples and is more suitable for 3D reconstruction under mass satellite remote sensing data. Comparative experiments on multiple datasets show that our method has a stereo matching EPE of 2.44 and a 3D reconstruction RMSE of 2.36 m. Especially in the weak texture and parallax abrupt change regions, we can achieve more advanced performance than other methods.

Keywords: satellite stereo images; self-supervised stereo matching; superpixel random walk; pre-matching; parallax-channel attention mechanism

Citation: Chen, W.; Chen, H.; Yang, S. Self-Supervised Stereo Matching Method Based on SRWP and PCAM for Urban Satellite Images. *Remote Sens.* **2022**, *14*, 1636. https://doi.org/10.3390/rs14071636

Academic Editors: Xiuping Jia, Chunhui Zhao, Wei Li, Shou Feng, Nan Su and Yiming Yan

Received: 14 February 2022
Accepted: 25 March 2022
Published: 29 March 2022

1. Introduction

In the past decade, automatic 3D reconstruction of urban scenes has been a research hotspot in remote sensing image processing, photogrammetry, and computer vision. However, 3D reconstruction from aerial images [1] and lidar point clouds [2] is difficult to scale to wider areas due to various constraints (e.g., air traffic control, airline authorization, equipment costs, etc.). In contrast, 3D reconstruction using optical satellite remote sensing stereo images is more advantageous in terms of data cost and coverage range [3].

Among the existing 3D reconstruction schemes of satellite data, most are based on the stereo correspondence of two-view images and use a series of processing steps such as rational function model to obtain 3D information of the scene. Among them, satellite image stereo matching, as a key step to realize the transition from 2D images to 3D models, has been a hot spot for research in the field of satellite 3D reconstruction. Although stereo matching of satellite images is oriented to more complex images, its essence is still to find the same name points from two images, just like stereo matching of ordinary images. According to the development trend of stereo matching, it can be divided into traditional methods and neural network methods.

However, both traditional methods and neural network methods still have much room for progress in the stereo matching of satellite images for urban scenes. The advantages and disadvantages of different methods are summarized as follows:

1. The traditional matching methods do not need a large number of training samples, which makes them faster and requiring less computational resources. They can obtain some high-confidence matching points in the local area of the image through simple and known feature artificial selection. However, in urban scenes, there are more complex situations on satellite images than in other scenes. Traditional methods are not as accurate as the CNN method. They only describe the matching cost using conventional features such as gradient, census, and scale-invariant feature transform (SIFT), which extract limited feature dimensions and make it difficult to achieve better results on satellite images [4].
2. Convolutional neural networks can extract deep features for finding more accurate stereo matching correspondence, which is more advantageous in processing larger amounts of remote sensing data. However, a large number of training samples containing truth labels are required for CNN training, which is difficult to obtain for stereo matching of satellite data [5]. If the network model trained with other datasets is directly used to match real satellite stereo remote sensing images, the effect is poor [6].

To this end, this paper uses traditional matching methods to obtain some matching points with high confidence as training samples, and convolutional neural networks to extract depth features for describing stereo matching relationships. Combining the advantages of traditional matching methods and convolutional neural networks, a self-supervised stereo matching method based on SRWP and PCAM is proposed for 3D reconstruction of satellite stereo images of urban scenes. Our method is divided into two stages: training and testing. First, in the training stage, we obtain pre-matching results by a pre-matching method based on superpixel random wandering, and select some of the matching points with high confidence as labeled training samples. Then, a stereo matching network is constructed to describe the correlation between two points by calculating the attention scores of any two points between different images through the parallax-channel attention mechanism, superimposing the attention scores of each layer to calculate the disparity. The network is trained on labeled samples and the unsupervised constraint criteria. Finally, in the testing phase, the trained network is used to re-match the stereo images to be matched. The main contributions of our work can be summarized as follows:

(1) A pre-matching method based on superpixel random walk is proposed. The occlusion and parallax discontinuity existing in stereo images are handled by constructing parallax consistency and mutation constraints. The matching cost update is achieved by superpixel segmentation and random walk to ensure the reliability of disparity for weak texture and parallax mutation regions. This method is robust to visual difference and occlusion between images with different viewing angles.
(2) A parallax-channel attention stereo matching network is proposed. The self-supervised training problem under sparse samples is solved by a feature enhancement module. The correspondence of stereo image pairs is captured by parallax-channel attention. The method can achieve better results for stereo matching of complex urban scenes.

The remainder of this paper is organized as follows. Section 3 provides a detailed description of the proposed method. Section 4 presents comparative experiments and a discussion of public data sets. Conclusions are drawn in Section 5.

2. Related Work

Traditional stereo matching methods can be broadly classified into local matching [7], semi-global matching (SGM) [8] and global matching [9]. The global matching usually optimizes a global objective function, while the local matching tends to consider the neighborhood information. The semi-global matching comprehensively considers the advantages of both. It essentially finds the optimal parallax for each pixel through dynamic programming, so that the global energy function of the whole image is minimized. The algorithm is insensitive to the effects of illumination changes and has strong robustness to noise [10]. Huang et al. [11] proposed an image-guided SGM method based on defining

the rules for propagation of valid pixels to invalid pixels by enhancing the propagation cost of texture-rich regions and guiding the interpolation method to interpolate the invalid parallaxes to obtain the final matching results, thus solving the problems of noise and ambiguity of similarity metrics in the matching process. Li et al. [12] proposed an improved SGM method by replacing the mutual information entropy in the original SGM with the census transform and reducing the matching time using a hierarchical matching strategy. Hosni et al. [13] proposed a fast cost volume filtering stereo matching (FCVFSM) method to realize a real-time disparity map by using fast edge preserving filtering to smooth label costs. In addition to this, some scholars have also started from other perspectives, such as Changjae et al. [14], who considered the matching cost as the probability of matching between points and sought the steady-state distribution of the matching probability through the proposed random walk with restart algorithm, thus giving the confidence of their matching along with the output matching result.

On the other hand, various data-driven convolutional neural network (CNN) methods have been introduced into the field of stereo matching, which can combine the disparity solution process into an end-to-end network by relying on massive training samples; these have gradually become the mainstream of stereo matching methods. Driven by the success of CNNs in many vision tasks, some early network-based methods used CNNs to replace one or more steps in the stereo matching process [15,16]. Zbontar et al. [17] predicted the matching degree of two image blocks by training a CNN, and then used the traditional cost aggregation method to calculate the matching cost to obtain the matching relationship. Kendall et al. [18] proposed GCNet, an end-to-end stereo matching network that obtains geometric and contextual information directly from binocular image data. It builds a matching cost cube with deep features, regularizes it through 3D convolution, and finally computes the best matching disparity value from the matching cost using a flexible argmin operation. Based on this, Chang et al. [19] proposed a pyramid pooling module for the feature extraction part of the GCNet, expanded the receptive field to obtain more representative features, and introduced a stacked hourglass network to regularize the matching cost cube. On the other hand, some scholars have tried to introduce traditional matching algorithms into the field of deep learning. Seki et al. [20] proposed a learning-based parameter estimation method, using CNNs to train and adjust the penalty parameters of SGM.

All the above stereo matching methods are designed for the stereo matching task of computer vision images; satellite remote sensing images are very different from computer vision images in terms of acquisition methods and image complexity. For example, images taken by satellites under high-speed motion conditions have longer baselines and larger angle differences than computer vision images, as the sides of inclined buildings are prone to inconsistent left and right views due to occlusion, their tops have a large number of weakly textured areas that are difficult to match, and their edges are difficult to locate accurately due to parallax abrupt change generated by differences in floor height [21].

For this reason, some scholars have begun to transfer various stereo matching methods to the stereo matching of satellite remote sensing images. Tatar et al. [22] and Mandanici et al. [23] directly applied SGM to the stereo matching of satellite images without modification, and achieved good matching results on roof surfaces and open spaces with weak textures in GeoEye-1 and WorldView-3 satellite data. Yang et al. [24] proposed a semi-global block matching (SGBM) to achieve disparity and height estimation in weakly textured waters by adaptive block matching. Zhu et al. [25] proposed a feature learning method based on a two-branch network to transform the image matching problem into a two-class classification problem, which used a two-stage training model to deal with the complex features of remote sensing images. Tao et al. [26] achieved stereo matching of remote sensing images using an improved pyramid stereo matching network and also improved the construction of matching cost to cope with the large variation of disparity range in remote sensing images, which can achieve better results in the parallax abrupt change region.

3. The Proposed Method

In this section, we present the proposed self-supervised stereo matching approach in detail, and the outline of our approach is given in Figure 1. Our self-supervised method is divided into two stages: training and testing. In the training stage, some labeled samples with high confidence are obtained through the proposed pre-matching method, which is then used to train the parallax-channel attention stereo matching network. After the net-work model is trained, the model is used to test the stereo images one by one in the testing stage to obtain the disparity map and stereo matching relationship. Therefore, the pre-matching method and the parallax-channel attention network during the training stage are the core of this paper and will be introduced in detail in this section.

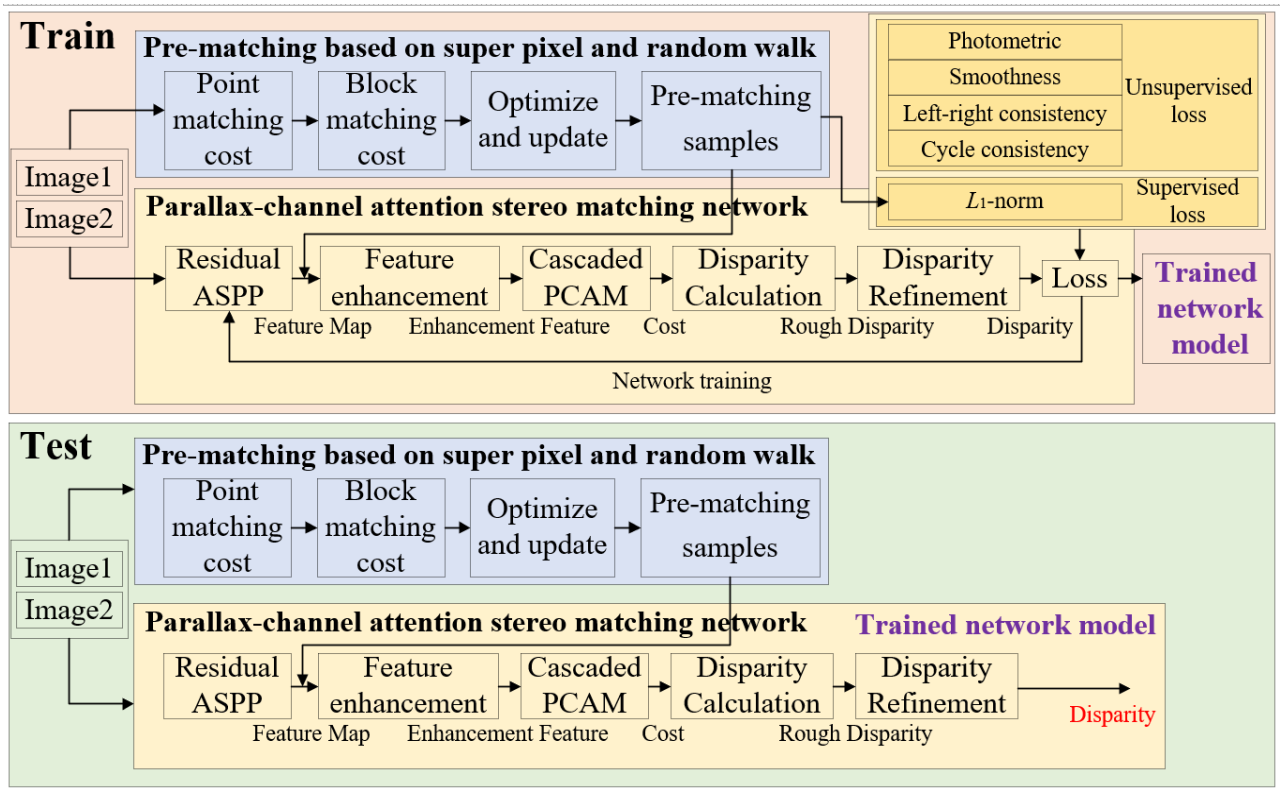


Figure 1. The training and testing process for proposed self-supervised stereo matching method based on SRWP and PCAM.

3.1. Superpixel Random Walk Pre-Matching

The main objective of this section is to take advantage of traditional matching methods to obtain some high-confidence matching points in local regions by means of artificially selected simple features. These matching points are then used as labeled training samples for subsequent matching networks. Generally speaking, sparse matching methods based on feature points and lines can obtain stable and accurate feature points for matching, but the number of matching points obtained by these methods is too small to support the training requirements of subsequent network models. In contrast, the dense matching method can obtain more matching points. The random walk algorithm proposed by Ham et al. [14] first converts the matching problem into a probability model. The matching cost is regarded as the probability of matching between points, which can provide a reference for us to screen the accuracy of pre-matching points. However, the inherent smoothing assumption of this method makes the matching unsatisfactory in areas of occlusion or parallax abrupt change. For this reason, we consider that the edges of the optical images of urban areas are semantically linked to the edges of the disparity map. Texture-similar regions of optical images present fixed or linear changes in disparity values on the disparity map,

without abrupt changes. In contrast, optical images with abrupt textures are prone to abrupt changes in disparity values. We can understand that there is a certain constraint relationship between the optical image and the disparity map, and since the disparity map is a form of representation of the stereo matching relationship, the use of optical image edge information can provide valuable clues to obtain the matching relationship between the parallax abrupt change region.

Based on the above ideas, this paper proposes a pre-matching method based on superpixel random walk, and the outline of the pre-matching is given in Figure 2. Using the constraint relationship between optical image and disparity image, the noise results in weak texture, and the mismatch of occlusion or parallax abrupt change region are removed by superpixel segmentation and two constraint criteria. Specifically, we first construct a point matching cost using selected simple features. Then, we aggregate them into block matching cost based on the superpixel segmentation results. Finally, the matching cost is updated and optimized according to the two constraint criteria of parallax consistency and mutation, so as to obtain some stable matching points. The method consists of three steps: constructing the point matching cost, constructing the block matching cost, and optimizing and updating the cost.

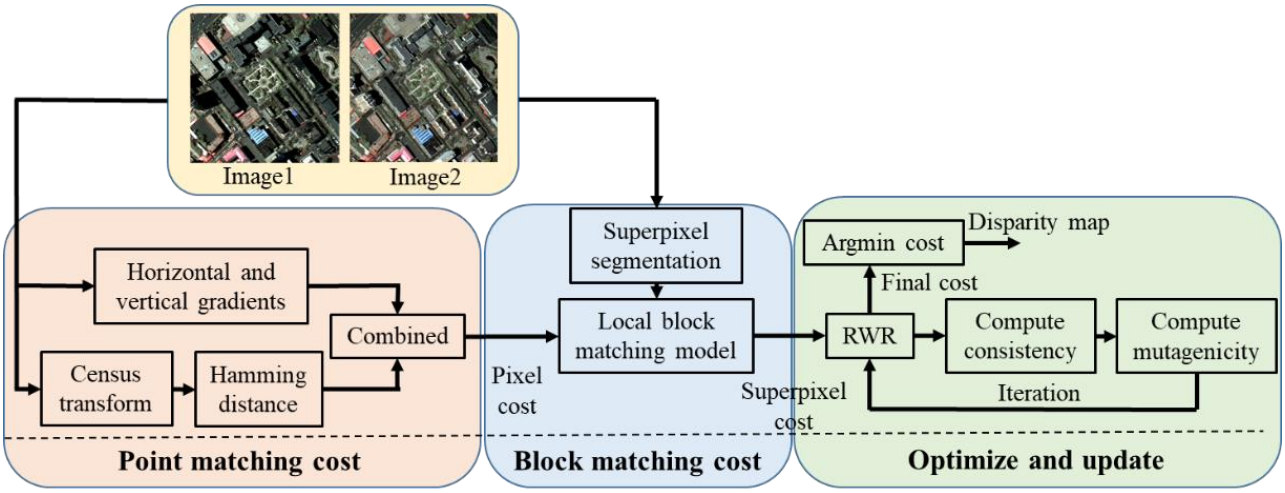


Figure 2. Outline of proposed pre-matching method based on superpixel and random walk.

3.1.1. Point Matching Cost

One of the most basic ideas of stereo matching is to describe the matching correlation between two images by constructing a matching cost function, so that the two points with the greatest correlation can be selected as matching points. In this stage, we first construct the initial matching cost of the pixel with common features in the existing matching methods. The gradient features, census transform, rank transform and mutual information features commonly used in stereo matching methods have high accuracy and stability. Compared with rank transformation and mutual information that requires initial disparity values and hierarchical iterations, this paper takes into account the computational needs of the subsequent block matching and optimization algorithms, as well as the fact that the urban scenes of interest contain a large number of buildings, so the initial matching cost of the pre-matching method is constructed using gradient features and census transform, which are more computationally efficient and more sensitive to building edges.

The census transform [27] technique is to convert the pixels of the left and right images into binary vectors and compare them to the surrounding pixels within finite support regions, as shown in Equation (1):

$$T(i,j) = \underset{(i_w,j_w)\in w(i,j)}{\oplus} H(I(i,j), I(i_w,j_w)) \quad (1)$$

where $I(i,j)$ and $I(i_w, j_w)$ denote the intensity values of the target pixel and the pixels around the target, respectively, $\oplus$ denotes the cascade, w is the window around (*i*,*j*), and H is the binary function that returns 0 or 1. We use a 5×5 window to encode a binary vector of each pixel in the Census transform. The binary vectors are encoded by comparing the intensity values of the center and its surrounding pixels, as in Equation (2):

$$H(I(i,j), I(i_w, j_w)) = \begin{cases} 0, \text{ if } I(i,j) < I(i_w, j_w) \\ 1, \text{ if } I(i,j) \geq I(i_w, j_w) \end{cases} \tag{2}$$

where $H(I(i,j), I(i_w, j_w))$ is the binary function of (i,j) and (i_w, j_w). The binary vector is assigned to each pixel in the left and right images. The matching cost is calculated using the Hamming distance [28] of the two binary vectors, as shown in Equation (3):

$$\begin{aligned} C_r(i,j,d) &= \text{Hamming}(T_l(i,j), T_r(i+d,j)) \\ C_l(i,j,d) &= \text{Hamming}(T_r(i,j), T_l(i-d,j)) \end{aligned} \tag{3}$$

where $C(i,j,d)$ is the matching cost based on Hamming distance at disparity d. The subscripts l and r denote the left image and right image, respectively. Since the census transform encodes the image structure based on the relative ordering of pixel intensities, it has better robustness to illumination variations and image noise. However, due to this property, matching blur may result in weakly textured areas with the same or similar textures. To solve these problems, we include gradient features in the calculation of the initial matching cost.

The matching cost based on image gradient features is defined as in Equation (4):

$$\begin{aligned} G_r(i,j,d) &= |\nabla_x I_l(i,j) - \nabla_x I_r(i+d,j)| + \left|\nabla_y I_l(i,j) - \nabla_y I_r(i+d,j)\right| \\ G_l(i,j,d) &= |\nabla_x I_r(i,j) - \nabla_x I_l(i-d,j)| + \left|\nabla_y I_r(i,j) - \nabla_y I_l(i-d,j)\right| \end{aligned} \tag{4}$$

where $G(i,j,d)$ is the matching cost based on gradient feature at disparity d. $\nabla_x I$ and $\nabla_y I$ denote the horizontal and vertical gradient images, respectively. The gradient images are calculated with a 5×5 sobel filter.

The census transform and gradient features are combined by weight to construct the following point matching cost, as shown in Equation (5):

$$\begin{aligned} P_r(i,j,d) &= \sigma_c \min(C_r(i,j,d), \tau_c) + \sigma_g \min(G_r(i,j,d), \tau_g) \\ P_l(i,j,d) &= \sigma_c \min(C_l(i,j,d), \tau_c) + \sigma_g \min(G_l(i,j,d), \tau_g) \end{aligned} \tag{5}$$

where σ_c and σ_g are the weight parameter to balance the census term and the gradient term, respectively. τ_c and τ_g are truncation values used to limit the influence of outliers. P_r is the matching cost of each pixel in the right image compared to each point on the epipolar line of the left image.

3.1.2. Block Matching Cost

The urban scenes contain a large number of artificial buildings. The most obvious feature of such buildings is the similarity texture of the building top surface, and the junction between buildings and non-buildings, which is prone to texture change. This property on the disparity map has similar performance. We use this property to aggregate the point matching cost into a block matching cost, so that there is a smooth parallax constraint within the block, while the inter-block is more prone to parallax abrupt change, as shown in Figure 3.

Superpixel is a block of images consisting of neighboring pixels with similar texture, color, and illumination characteristics. Different pixels in one superpixel may have the same geometric features and similar parallax. Thus, segmenting the optical image by superpixels has similar results to segmenting the parallax map. For this reason, we use each superpixel block segmented from the optical image as a guide for our aggregation point matching

cost. In this paper, we use the simple linear iterative clustering [29] to perform superpixel segmentation on the left and right images.

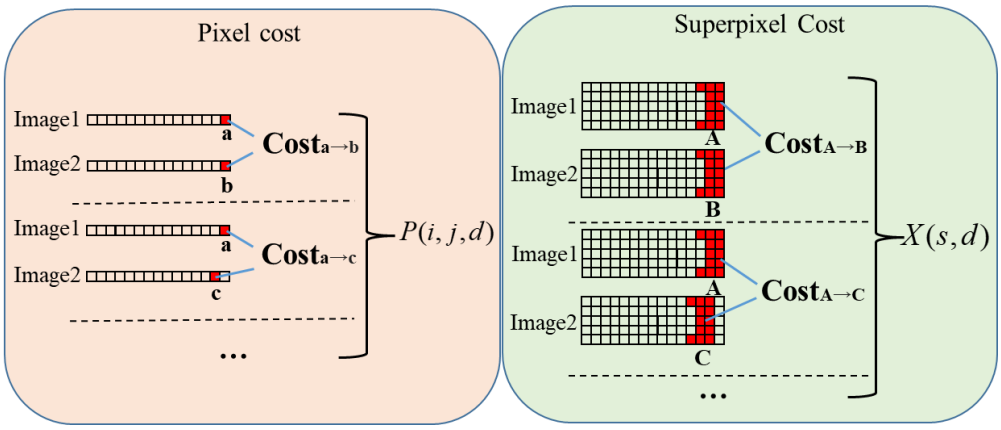


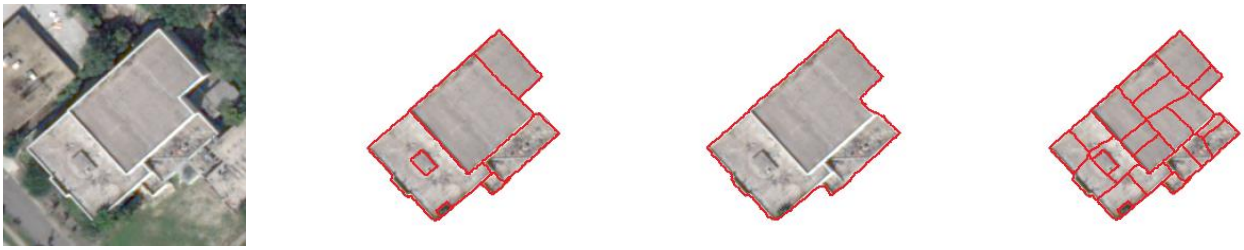
Figure 3. Comparison of pixel cost and superpixel cost.

The block matching cost can be given by Equation (6):

$$X(s,d) = \frac{1}{n_s} \sum_{(i,j)\in s} P(i,j,d) \tag{6}$$

where s is a superpixel block, $X(s,d)$ is the cost function of the superpixel s when the disparity is d, and n_s is the number of points in the superpixel s. $P(i,j,d)$ represents the point matching cost when the disparity is d at (i,j) in the superpixel s. The left image matching cost $X_l(s,d)$ and the right image matching cost $X_r(s,d)$ are calculated separately.

Although we construct the matching cost function for local blocks, the segmentation results of the superpixel will largely affect the matching results. Specifically, there are two problems: the larger the superpixel block, the more likely it is to have under-segmentation, where regions with different disparity are segmented in one superpixel block; the smaller the superpixel block, the more likely it is to have over-segmentation, where regions with the same disparity are segmented in different superpixel blocks, and as shown in Figure 4.



(**a**) (**b**) (**c**) (**d**)

Figure 4. Results of different superpixel segmentation cases: (**a**) optical image; (**b**) accurate segmentation; (**c**) under-segmentation; (**d**) over-segmentation.

3.1.3. Optimization and Updating

In the actual segmentation process, it is difficult to guarantee that accurate segmentation results can be obtained every time, which causes a certain error in the block matching cost. For this reason, in this paper, the matching cost of each superpixel block is updated iteratively under the condition of considering the influence of surrounding blocks, provided that the superpixel chunks are small enough, so as to achieve a stable block matching cost. This idea of iteratively updating the matching cost is similar to the random walk algorithm; both of them are designed to obtain a smooth stable probability distribution or matching cost. To this end, this paper improves the random walk algorithm to update the block

matching cost and eliminate the interference caused by over-segmentation through constraints such as smoothness, consistency and mutability. The final result of block matching cost aggregation is similar to that of superpixel exact segmentation.

Random walk was first proposed for image segmentation [30]. It starts from a node in the graph and faces two choices at each step, randomly choosing an adjacent node or returning to the starting node. The algorithm contains a parameter c for the restart probability and $1 - c$ for the probability of moving to an adjacent node. After iterations to reach stability, this probability distribution can be considered as the distribution influenced by the start node. We apply the random walk to the block matching cost update, and the update function is defined as Equation (7):

$$X_{t+1}^d = c\overline{W}X_t^d + (1-c)X_0^d \tag{7}$$

where $X_0^d = [F(s,d)]_{k\times 1}$ denotes the initial matching cost when the disparity value is d, X_t^d denotes the updated matching cost, t is the number of iterations, and k is the number of superpixels. The weighting matrix $W = [w_{uv})]_{k\times k}$ contains the edge weights of all superpixels, and $\overline{W}$ is obtained by normalizing the rows of W. Edge weights are used to describe the probability that the matching cost of a superpixel block is passed to the neighboring blocks. We assume that neighboring superpixel blocks on an optical image tend to have similar disparity values on the disparity map when the color distance is close. Therefore, neighboring superpixels with similar intensities have more influence on each other. The edge weight w_{uv} of the u-th and v-th superpixel blocks is calculated by Equation (8):

$$w_{uv} = (1-\tau_e)e^{-\frac{(I(s_u)-I(s_v))^2}{\sigma_e}} + \tau_e \tag{8}$$

where $I(s_u)$ and $I(s_v)$ are the intensities of the u-th and v-th superpixel blocks, respectively, and τ_e and σ_e are parameters that control the shape of the function.

The matching cost X gradually reaches convergence as the number of iterations t increases. The above method provides a local minimum, but the limitations of the smoothness constraint mean that it does not provide a good solution in regions of occlusion or parallax abrupt change. Therefore, we added parallax consistency and mutation constraints to correct and optimize the matching cost for these regions.

The occluded pixels involved in this paper are the pixel points that appear in only one view and are not visible in the other view. In order to eliminate the effect of occluded pixels on the matching cost update, we use parallax consistency to detect occluded pixel blocks and set the occluded pixel blocks to zero in the matching cost update process. Parallax consistency means that the matching relationships obtained in the two views should correspond to each other, and the occluded pixels do not satisfy this consistency. Therefore, we propose the following consistency constraint function, as shown in Equation (9):

$$O_t(s) = \begin{cases} 1, \text{ if } |D_r(x_s,y_s) - D_l(x_s + D_r(x_s,y_s),y_s)| \le 1 \\ 0, \text{ if } |D_r(x_s,y_s) - D_l(x_s + D_r(x_s,y_s),y_s)| > 1 \end{cases} \tag{9}$$

where D_l and D_r are the current parallax maps of the left image and right image, respectively, and x_s and y_s are the x and y centroids of superpixel s. The superpixel blocks with inconsistent disparity in the left and right disparity maps are divided into occluded superpixels and set to 0, while the other blocks are set to 1 as non-occluded superpixels. The occlusion masks $V_t = [O_t(s)]_{k\times 1}$ are obtained by splicing each $O_t(s)$. Finally, the matching cost is multiplied by the occlusion mask to obtain the consistent matching cost after the parallax consistency constraint, as shown in Equation (10):

$$v_t^d = X_t^d \Theta V_t \tag{10}$$

where Θ denotes the element-wise product function.

The random walk algorithm considers that adjacent blocks have more influence on each other, which is manifested in the parallax map by the existence of smoothness constraints and prone to errors in the parallax abrupt change region. For example, in the eaves of a building, the disparity value varies greatly, but the disparity boundary becomes blurred due to the smoothness constraint. To prevent such problems, we add a mutability constraint. First, we calculate the temporary disparity value of the superpixel based on the current matching cost, as shown in Equation (11):

$$d'_u = \frac{\sum\limits_{v \in N(u) \cup u} w_{uv} \overline{d}_v O_t(s_v)}{\sum\limits_{v \in N(u) \cup u} w_{uv} O_t(s_v)} \tag{11}$$

where w_{uv} is the edge weight, $O_t(s_v)$ is the consistency constraint, $\overline{d}_v$ is the current disparity of the neighboring superpixel, and d'_u is the temporary disparity value of the u-th superpixel. The mutability matching cost is calculated using the temporary disparity values composed of all superpixel blocks, as in Equation (12).

$$\psi_t^d(d') = \begin{cases} \left((d' - d)/\sigma_\psi\right)^2, & \text{if } |d' - d| \le \tau_\psi \\ \left(\tau_\psi/\sigma_\psi\right)^2, & \text{if } |d' - d| > \tau_\psi \end{cases} \tag{12}$$

where d' is the Equation (11) calculated parallax, σ_ψ is the scalar parameter, and τ_ψ denotes the truncation parameter, which play an important role in controlling the parallax mutability.

The mutability constraint preserves disparity boundaries by maintaining the intensity difference between adjacent superpixels, avoiding blurring small objects into the background and thus preserving more detailed information.

Combining parallax consistency and mutability matching cost, we construct the following block matching cost iterative update function, as shown in Equation (13):

$$X_{t+1}^d = c\overline{W}\left((1-\lambda)V_t^d + \lambda\Psi_t^d\right) + (1-c)X_0^d \tag{13}$$

where Ψ_t^d is the mutability matching cost calculated in Equation (12), V_t^d is the consistency matching cost calculated according to Equation (10), λ is used to balance them, and c is the restart probability. The consistency and mutability matching costs are determined based on the current matching cost X_t^d. The matching cost propagates along the graph $\overline{W}$, and the initial matching cost is aggregated into the current matching cost, which is proportional to the restart probability $(1 - c)$. The combination of the superpixel matching cost and the initial point matching cost constitutes the final matching cost P, and the parallax value $\hat{d}$ is determined by minimizing the matching cost, as in Equation (14):

$$\begin{gathered} P_{t+1}(i,j,d) = X_t^d(s,d) + \gamma P_t(i,j,d) \\ \hat{d} = \underset{d}{\operatorname{argmin}}(P(i,j,d)) \end{gathered} \tag{14}$$

where s is the superpixel corresponding to pixel (i,j), γ denotes the weight of the superpixel and the point matching cost, and argmin means finding a disparity value to minimize the matching cost P.

Since the matching cost P describes the degree of matching between two image point pairs, we can filter all the matching points by setting a threshold to obtain some of the pairs with higher matching confidence, which we call pre-matched pairs.

3.2. Stereo Matching Network Based on Parallax-Channel Attention Mechanism

Although the rough matching relationship of images can be obtained using the pre-matching method, conventional features such as gradient, census, and SIFT have limited feature dimensionality for extraction. Through the combination and iterative update, even if the artificial buildings and other areas we are concerned about can have a good matching effect, there is a significant decrease in accuracy for other areas of the city. To solve this problem, this paper first constructs a stereo matching network to describe the stereo matching relationship using the deep features extracted by the CNN. In order to further improve the network performance, we use the high-confidence pre-matched point pairs obtained in the previous section and some unsupervised criterion to supervise and constrain the network. Finally, a stereo matching network supporting sparse training samples is formed.

In recent years, the self-attention mechanism [31] has been used in various fields of visual image processing based on CNNs. The self-attention mechanism forms an attention score by calculating the similarity of extracted features at any two points in a single image, and then weights this score to the image to highlight more significant parts. We note that the self-attention mechanism calculates the similarity of two points of an image as the attention score, and the purpose of stereo matching is to find two points that are similar in different images. Therefore, this paper introduces the process of obtaining attention scores from the self-attention mechanism into stereo matching and describes the matching correlation of two points by calculating the attention scores of any two points between different images. Since a single attention score map can depict only one disparity value, we form a 3D attention map by stacking the attention scores of different disparity values. The i-th layer in the 3D attention map represents the attention score of each pixel with disparity i. Finally, the disparity is calculated by superimposing the attention score maps of each layer to achieve stereo matching.

With the above ideas, this paper constructs a parallax-channel attention mechanism stereo matching network based on the traditional parallax attention for image super-resolution and channel attention for target detection. Figure 5 shows the structure of the proposed stereo matching network. We first extract the deep features by hourglass network, then use the cascaded parallax-channel attention module to calculate the extracted features as attention scores and combine them into 3D attention maps, and finally calculate the disparity by superimposing the attention maps of each layer. In addition, we add two optimization strategies, namely feature enhancement and disparity refinement, which are used to obtain more significant features in feature extraction and more refined disparity in disparity calculation.

Recent CNN matching methods, such as GCNet [18], PSMNet [19], etc., usually focus on feature extraction of higher dimensions to generate larger 4D cost volumes (height $\times$ width $\times$ maximum disparity $\times$ feature dimension), and then use 3D CNNs to achieve disparity map acquisition. However, such methods require high computational and memory costs and therefore require manual setting of the maximum disparity to ensure processing efficiency, which in turn limits their ability to process urban images with large disparity variations. In addition, these methods have more complex network structures, contain more dimensions in the 4D cost volume, require larger training samples, and are difficult to adapt to the practical requirements of self-supervised matching for satellite remote sensing images. The stereo matching network proposed in this paper has low computational and memory costs, and ensures the possibility of matching between any two points of different views on the epipolar line without setting the maximum disparity. Relying on multiple unsupervised constraints ensures that good matching results are achieved even with only a small number of labeled samples obtained by pre-matching.

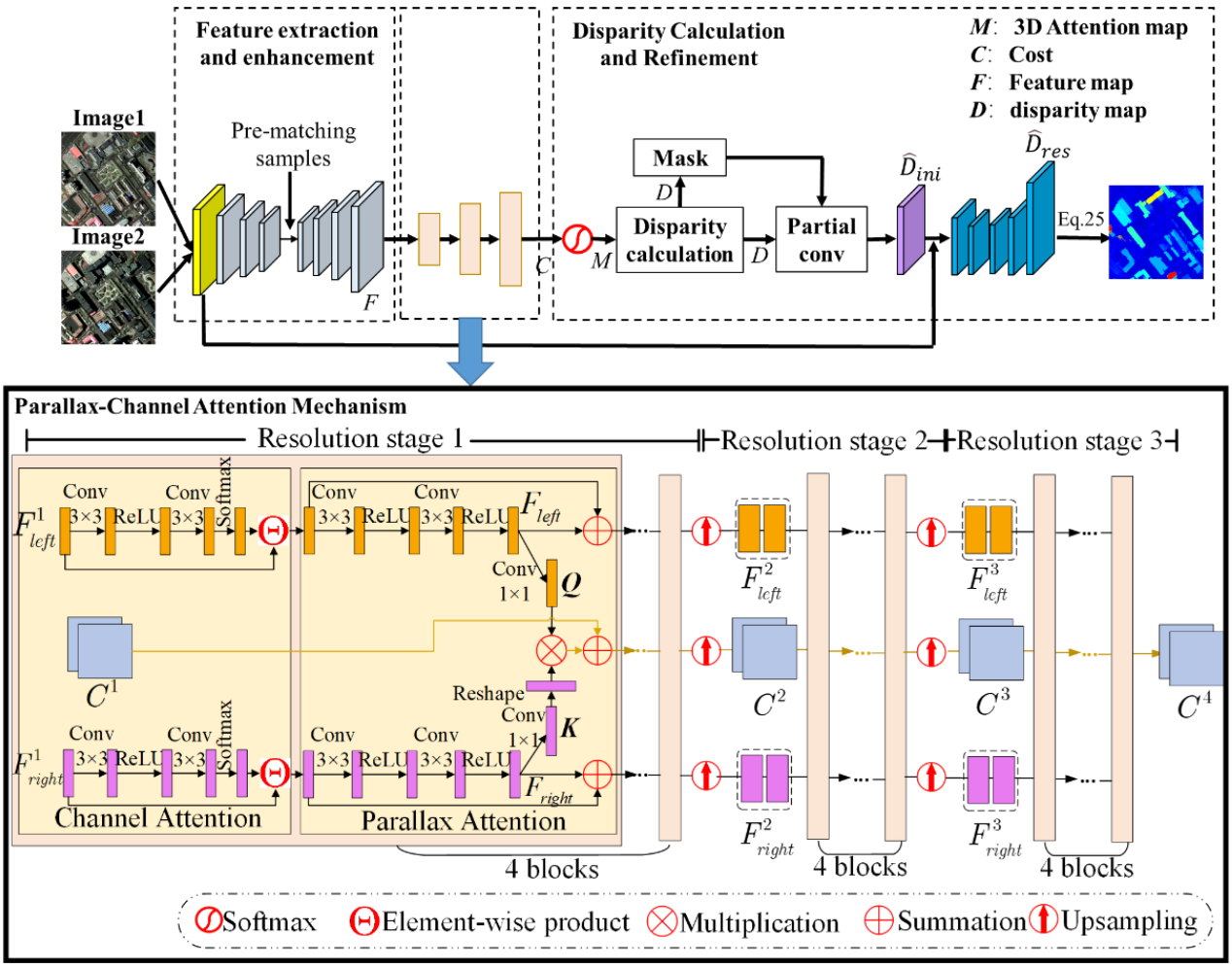


Figure 5. Architecture overview of proposed stereo matching network based on parallax-channel attention mechanism.

3.2.1. Feature Extraction and Enhancement

The partially labeled samples obtained by pre-matching can be applied to the stereo matching network as training samples. However, considering that our proposed self-supervised network uses the same data in both the training and testing stage, and that a batch of pre-matched point pairs with high confidence is available for all these data, we can use this prior information to obtain more discriminative features in the feature extraction. In this paper, we propose a feature enhancement process in the feature extraction stage, using the labeled samples obtained by previous pre-matching to enhance the extracted features at these samples by Gaussian functions, so that the points at the labeled samples can obtain higher attention scores on the subsequent attention map. The essential idea is to give more weight to the features related to the labeled samples obtained by pre-matching on the feature map, so as to ensure that the pre-matching points with high confidence retain the original matching results as much as possible.

First, we use the residual atrous spatial pyramid pooling module [32] for feature extraction of left and right images to obtain hierarchical features with dense pixel sampling rate and scale. After that, we expand the labeled samples obtained from the previous pre-matching into two new inputs, a sparse matrix D_{gt} of size $H \times W$ to represent the sparse pre-matching disparity map, and another binary mask D_{loc} of the same size for the specified D_{gt} locations that have pre-matching samples. For each pixel in the image with position (*i*,*j*) and $D_{loc}(i,j) = 1$, we enhance the features by using a Gaussian function centered on $D_{gt}(i,j)$. The feature corresponding to $D_{loc}(i,j) = 1$ is multiplied by the peak

value of the function, while any other pixels are gradually multiplied by a lower factor until they become far away from (*i*,*j*) and are suppressed. Specifically, our enhancement function U is defined as in Equation (15):

$$U = k \cdot e^{-\frac{dis^2}{2\sigma^2}} \tag{15}$$

where σ determines the width of the Gaussian curve, k denotes its maximum amplitude and shall be greater than or equal to 1, and dis denotes the distance between the current pixel point and (*i*,*j*). Therefore, we obtain the new enhanced feature F by multiplying the original feature F_{ini}, and we construct the following feature enhancement function as shown in Equation (16):

$$F = (1 - D_{loc} + D_{loc} \cdot U) \cdot F_{ini} \tag{16}$$

where the weight factor on the left is equal to 1 when $D_{loc} = 0$.

3.2.2. Parallax-Channel Attention

Parallax Attention

The parallax attention mechanism was first proposed to improve the performance of stereo image super-resolution with a global field of perception along the epipolar line to process different stereo images with large parallax variations [33]. Similar to the self-attention mechanism, parallax attention describes the matching correlation between two points by computing the attention score of any two points between different images and then weighting this score to the image to achieve super-resolution. We achieve stereo matching with this attention mechanism by calculating the correlation between pixels in the left and right image through the dot product of features and using it as the matching cost. Then, using softmax to achieve normalization, we generate the attention score and combine it into a 3D attention map. Finally, we calculate the disparity in the next section to achieve stereo matching, as shown in Figure 6.

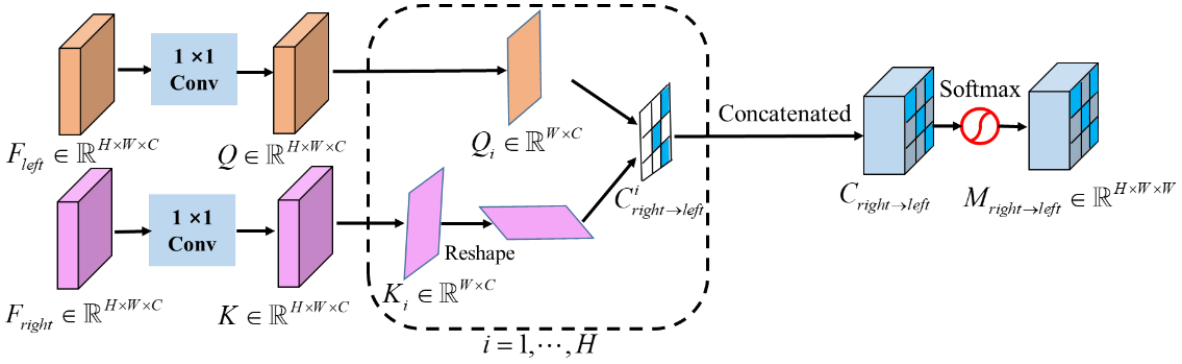


Figure 6. An illustration of parallax attention.

Specifically, the corresponding feature maps $F_{left}, F_{right} \in \mathbb{R}^{H\times W\times C}$ are extracted from the stereo images, where H and W are the height and width of the image, respectively, and C is the dimension of the extracted deep features. The F_{left} is fed to a 1×1 convolution to produce a query feature map $Q \in \mathbb{R}^{H\times W\times C}$, the F_{right} is fed to another 1×1 convolution to produce a key feature map $K \in \mathbb{R}^{H\times W\times C}$, and then the i-th row $Q_i \in \mathbb{R}^{W\times C}$ and $K_i \in \mathbb{R}^{W\times C}$ in Q and K are extracted, respectively. Then, we perform matrix multiplication between Q_i and $K_i{}^T$ and concatenate the matrix multiplication results from 1-th to H-th row to obtain the matching cost $C_{right\to left} \in \mathbb{R}^{H\times W\times W}$. Finally, the 3D attention map $M_{right\to left} \in \mathbb{R}^{H\times W\times W}$ is generated by softmax normalization.

Channel Attention

The image features extracted using residual atrous spatial pyramid pooling contain multiple channels. Parallax attention only considers the spatial features of the image features, equivalently treating each channel of the features and ignoring the differences between the different channels. For this reason, we use a channel attention mechanism [34] that applies feature weighting to the different channels of the extracted features F_{left} and F_{right}. It models the importance of each feature channel and then enhances or suppresses different channels, so that the channels that are more favorable for matching accuracy improvement receive more attention and further improve the matching effect. Since each channel describes different semantic information of the image, applying the attention mechanism in the channel can be regarded as a process of selecting semantic attributes. The illustration of channel attention mechanism is given in Figure 7.

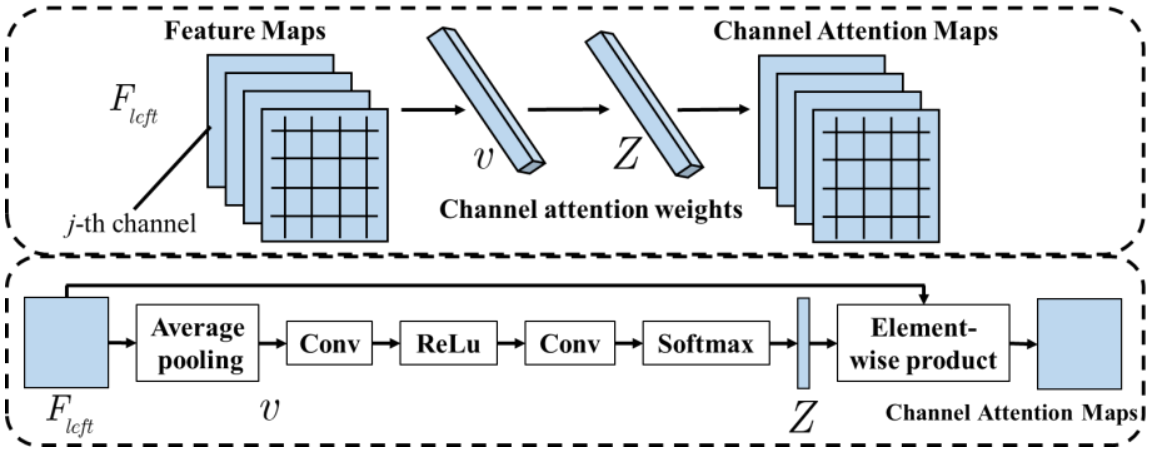


Figure 7. An illustration of channel attention.

Take the feature map $F_{left} \in \mathbb{R}^{H \times W \times C}$ as an example, which is obtained from the left image. The feature map F_{left} is first reshaped to $F_{left} = [f_1, f_2, \cdots, f_C]$, which has C feature maps with a size of $H \times W$. Then the channel statistics feature v_k of the k-th channel can be obtained by Equation (17):

$$v_k = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} f_k(i, j) \tag{17}$$

where $f_k(i, j)$ is the feature value of the k-th layer feature map at the position (i, j). The above function can be interpreted as average pooling. The average pooling is then applied to each channel to obtain the channel features $v = [v_1, v_2, \cdots, v_C]$ and calculate the channel attention weights Z as shown in Equation (18):

$$Z = \delta(W_U R(W_D v)) \tag{18}$$

where $R()$ denotes the leaky ReLU function, $\delta()$ represents the softmax function, W_D refers to the channel-downscaling convolution layer, and W_U refers to the channel-upscaling convolution layer.

The obtained channel attention weights Z are then used to rescale the feature maps F in the following manner to obtain channel attention weighted feature maps F', as shown in Equation (19):

$$F' = F \Theta Z \tag{19}$$

where Θ denotes the element-wise product function.

Cascaded Parallax-Channel Attention Module

The matching network constituted by the above two attention mechanisms is still essentially a local matching. It only considers the local information and lacks the utilization of the global information. It is necessary to expand the support window of matching cost, realize matching cost aggregation, reduce the influence of anomalies, improve signal-to-noise ratio and improve matching accuracy [35]. To this end, this paper achieves matching cost aggregation by cascading parallax-channel attention modules of different resolutions and continuously superimposing the matching costs of each level to form the final matching cost, while expanding the receptive field. Our cascaded parallax-channel attention module consists of three resolution stages, each of which consists of four modules in series, as shown in Figure 5.

First, the extracted features $F^1_{left}, F^1_{right} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$ are input at the first resolution stage, and the initial matching cost $C^1_{right\to left}, C^1_{right\to left} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times\frac{W}{16}}$ has a construction value of 0. The F^1_{left} and F^1_{right} are fed into the channel attention module to obtain the channel weighted feature maps ${F^1_{left}}'$ and ${F^1_{right}}'$, respectively. Then, the F_{left} and F_{right} are obtained by a 3×3 convolution shared by two parameters. Next, the query feature Q and the key feature K are obtained from F_{left} and F_{right} by 1×1 convolution, and K is reshaped and multiplied with Q to obtain the matching cost $C_{right\to left}$. Once $C_{right\to left}$ is ready, F_{left} and F_{right} are exchanged to obtain $C_{left\to right}$. After that, $C_{right\to left}$ and $C_{left\to right}$ are added to $C^1_{right\to left}$ and $C^1_{left\to right}$, and F_{left} and F_{right} are added to F^1_{left} and F^1_{right}. After four repetitions of this module and bilinear upsampling, the input of the next stage $C^2_{right\to left}$, $C^2_{left\to right}$, F^2_{left} and F^2_{right} are obtained. The above steps are repeated at higher resolution stages until the final matching cost $C^4_{right\to left}, C^4_{left\to right} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times\frac{W}{4}}$ is obtained.

3.2.3. Disparity Calculation and Refinement

As shown in Figure 5, the final matching cost $C^4_{right\to left}$ sum $C^4_{left\to right}$ is obtained using the above cascade module. It is normalized by softmax, and the matching cost is converted into an attention score (between 0 and 1) to obtain a 3D attention map $M^4_{right\to left}, M^4_{left\to right} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times\frac{W}{4}}$. Disparity is calculated by superimposing the attention maps of each layer, as in Equation (20):

$$\hat{D} = \sum_{k=0}^{W/4-1} k \times M^4_{right\to left}(:,:,k) \tag{20}$$

So far we have obtained a rough parallax map at low resolution. To further obtain a more accurate parallax map, we achieve disparity refinement by occlusion handling and upsampling.

We obtain the occlusion mask using Equation (21), based on the same idea of processing the occlusion superpixel block in Section 3.1.3.

$$V(i,j) = \begin{cases} 1, \text{ if } \left|\hat{D}_r(i,j) - \hat{D}_l(i+\hat{D}_r(i,j),j)\right| \le 1 \\ 0, \text{ if } \left|\hat{D}_r(i,j) - \hat{D}_l(i+\hat{D}_r(i,j),j)\right| > 1 \end{cases} \tag{21}$$

where $\hat{D}_l$ and $\hat{D}_r$ are the current disparity maps of the left and right images, respectively.

Parallax is not available for the occluded region because the correspondence cannot be found for the pixels. However, in order to reduce the error on the occluded area during metric evaluation, we use an image patching method to approximately fill the disparity value of the occluded area by partial convolution [36].

Finally, we use a disparity upsampling module that uses the features of the left image as a guide to provide structural information such as edges [37]. As shown in Figure 5, the initial disparity $\hat{D}_{ini}$ and features F^4_{left} are concatenated in series and fed to the hourglass

network to produce a residual disparity map $\hat{D}_{res}$ and confidence map M_{con}. Finally, the refined disparity is calculated as Equation (22):

$$D = (1 - M_{con}) \times \hat{D}_{ini} \uparrow + M_{con} \times \hat{D}_{res} \tag{22}$$

where $\uparrow$ is a bilinear upsampling operator.

3.2.4. Losses

In this paper, we first consider stereo matching as a special case of optical flow estimation [38], and we use the photometric and smoothness losses in the optical flow framework as unsupervised loss constraints. After that, we define two additional unsupervised losses based on the left–right consistency and circular consistency of the stereo image. Finally we use the labeled samples obtained by pre-matching and employ the L_1 loss as a supervised loss to further optimize the disparity map generated by the network. We define the total loss as in Equation (23):

$$L = \lambda_p L_p + \lambda_s L_s + \lambda_{PCAM} L_{PCAM} + L_l \tag{23}$$

where λ_p, λ_s and λ_{PCAM} are used to balance different losses.

1. Photometric Loss

The photometric loss consists of a structural similarity (SSIM) index loss term and a mean absolute error (MAE) loss term. Where SSIM is used to measure the similarity of two images (i.e., one view and the other view after parallax transformation) and MAE is used to measure the mean mode length of the predicted value error [39]. Since photometric consistency only holds in non-occluded regions, the photometric loss is defined as Equation (24):

$$\begin{gathered} L_p = \frac{1}{N} \sum_{p \in V_{left}} \alpha \frac{1 - \mathcal{S}(I_{left}(p), {I_{left}}'(p))}{2} + (1 - \alpha) \| I_{left}(p) - {I_{left}}'(p) \|_1 \\ {I_{left}}' = I_{right} \odot D \end{gathered} \tag{24}$$

where $\mathcal{S}$ is a SSIM function, p indicates the non-occluded effective pixels, N is the number of effective pixels, and α is used to balance SSIM and MAE. $\odot$ is a warping operator using the disparity, which indicates that the pixels of the right view I_{right} are rearranged into $\hat{I}$ according to disparity D_{right}. The specific process is that each pixel point I_{right} in the image first find the disparity value d at the same position in the disparity map D_{right}, and then each pixel point of I_{right} is shifted d pixels to the right or left to obtain a new reconstructed image ${I_{left}}'$.

2. Smoothness Loss

We use an edge-aware smoothness loss to encourage local smoothness of the disparity [40,41], which is defined as in Equation (25):

$$L_s = \frac{1}{N} \sum_p (\| \nabla_x D(p) \|_1 e^{-\| \nabla_x I_{left}(p) \|_1} + \| \nabla_y D(p) \|_1 e^{-\| \nabla_y I_{left}(p) \|_1}) \tag{25}$$

where ∇_x and ∇_y are the gradients in the x and y axes, respectively.

3. PCAM Loss

Considering that the stereo image observes the same ground object from different angles, we define two additional unsupervised losses, left–right consistency loss L^k_{PCAM-q} and cyclic consistency L^k_{PCAM-c} loss, which are applied to adjust the PCAM stereo matching

network on multiple resolution scales. The PCAM loss terms for different resolution scales k (k = 1, 2, 3) are defined as in Equation (26):

$$L_{PCAM}^{k} = \lambda_{PCAM-q} L_{PCAM-q}^{k} + \lambda_{PCAM-c} L_{PCAM-c}^{k} \tag{26}$$

- L_{PCAM-q}^{k}

Left–right consistency means that the optical image of a viewing angle can move each pixel of the optical image according to the disparity value of the viewing angle as a guide for pixel movement, so as to obtain a reconstructed optical image. The reconstructed image is very similar to the optical image from another viewing angle. Thus, the more accurate the disparity value, the more similar the reconstructed image is to the optical image of the other viewpoint. The following function expresses the left–right consistency of stereo images, as shown in Equation (27):

$$\begin{cases} {I_{left}}' = I_{right} \odot D_{right} \\ {I_{right}}' = I_{left} \odot D_{left} \end{cases} \tag{27}$$

We use the left–right consistency to connect the relationship between the left and right optical images by modifying the above function to define the left–right consistency loss, as shown in Equation (28):

$$\begin{aligned} L_{PCAM-q}^{k} &= \frac{1}{N_{left}^{k}} \sum_{p \in V_{left}^{k}} \left\| I_{left}^{k}(p) - (I_{right}^{k} \odot D_{right}^{k})(p) \right\|_1 \\ &+ \frac{1}{N_{right}^{k}} \sum_{p \in V_{right}^{k}} \left\| I_{right}^{k}(p) - (I_{left}^{k} \odot D_{left}^{k})(p) \right\|_1 \end{aligned} \tag{28}$$

where N_{left}^{k} and N_{right}^{k} are the effective number of pixels in V_{left}^{k} and V_{right}^{k}, respectively. I_{left}^{k} and I_{right}^{k} are the bilinear downsampled images of the corresponding scale levels.

- L_{PCAM-c}^{k}

Cyclic consistency is based on the left–right consistency by reconstructing the reconstructed optical image according to the disparity value of another viewing angle. The following function represents the cyclic consistency of stereo images, as shown in Equation (29):

$$\begin{cases} {I_{right}}' = I_{right} \odot D_{right} \odot D_{left} \\ {I_{left}}' = I_{left} \odot D_{left} \odot D_{right} \end{cases} \tag{29}$$

We use cyclic consistency to further connect the relationship between the left and right optical images by modifying the above function to define cyclic consistency loss, as shown in Equation (30):

$$\begin{aligned} L_{PCAM-c}^{k} &= \frac{1}{N_{left}^{k}} \sum_{p \in V_{left}^{k}} \left\| I_{left}^{k}(p) - (I_{left}^{k} \odot D_{left}^{k} \odot D_{right}^{k})(p) \right\|_1 \\ &+ \frac{1}{N_{right}^{k}} \sum_{p \in V_{right}^{k}} \left\| I_{right}^{k}(p) - (I_{right}^{k} \odot D_{right}^{k} \odot D_{left}^{k})(p) \right\|_1 \end{aligned} \tag{30}$$

4. L_1 loss

L_1 loss is widely used for various tasks in image processing due to its robustness and low sensitivity to outliers [42]. We use the L_1 norm of the predicted disparity value obtained by the network and the disparity value obtained by pre-matching as the loss constraint. The L_1 loss is defined as in Equation (31):

$$L_l = \frac{1}{N} \sum_{p \in V_{left}} L(D_{pred}(p) - D_{gt}(p)) \tag{31}$$

in which

$$\mathcal{L}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{32}$$

where $D_{pred}(p)$ is the disparity of pixel p obtained through the network, and $D_{gt}(p)$ is the pre-matching disparity of pixel p.

4. Experimental Results and Discussion

4.1. Data Sets, Metrics, and Implementation Details

- Dataset

The datasets used in this paper comprise three sets. Dataset A is a set of public datasets from the IGRASS Data Fusion Contest in 2019. The source data includes images collected through worldview-2 and worldview-3 in Jacksonville and Omaha from 2014 to 2016, through cutting and other means, with a stereo image composed of 1685 pairs finally formed. The size of each image is 1024 × 1024 pixels, and ground sample distance (GSD) is 0.35 m. Dataset A provides the true disparity value corresponding to the left viewing angle at the same time. Dataset B is the selected 100 sets of stereo pairs in Dataset A, with manual markings of the area with weak texture or parallax abrupt change region as a mask. Dataset C comprises SuperView-1 satellite stereo images and digital ground model (DSM) data in Harbin; the image GSD is 0.5 m, and the digital ground model is measured by airborne LiDAR.

- Metrics

The most direct way to evaluate the performance of stereo matching is to compare disparity maps. However, considering the difficulty of obtaining the true value of disparity, there are many studies that use the rational function model to solve the matching results into DSM and indirectly evaluate the matching performance by comparing DSM. In this paper, disparity map comparison and DSM comparison will be used to evaluate the matching performance of this method in Datasets A, B and C.

For the evaluation of disparity results, we use end-point error (*EPE*) and the fraction of erroneous pixels (D1 and D3) as measures. The smaller the value of the three indicators, the better the matching effect. D1 represents the proportion of pixels with a real disparity error of more than 1 pixel, D3 represents the proportion of pixels with a real disparity error of more than 3 pixels, and *EPE* represents the average error between all pixel disparity and real disparity. The calculation function is defined as in Equation (33):

$$EPE = \frac{1}{m}\sum_{i=1}^{m}|d_i - dr_i| \tag{33}$$

where m is the number of valid pixels in the true disparity, d_i represents the disparity obtained by this method, and dr_i represents the true disparity.

We use the root mean square error (*RMSE*) to evaluate the DSM. The calculation function is defined as in Equation (34):

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(z_i - h_i)^2} \tag{34}$$

where z is the solved altitude and h is the altitude measured by LiDAR as the true value of the evaluation.

The above evaluation is only for points with true disparity or altitude.

- Implementation Details

The experiment was performed on a PC with Intel Core i7-10870H CPU, 16 G RAM, and Nvidia RTX 2080 GPU, and we did not use any parallel programs or other dedicated hardware. The PCAM network architecture was implemented using PyTorch. All models

were optimized using the Adam method [43], with β1 = 0.9, β2 = 0.999 and a batch size of 32. The initial learning rate was set to 1×10^{-4} for 60 epochs and decreased to 1×10^{-5} for another 20 epochs.

4.2. Evaluation of Pre-Matching Performance

To verify the effectiveness of the proposed pre-matching method, we first conducted experiments on the method proposed in Section 3.1. Considering that the purpose of pre-matching is to obtain stable and accurate matching points, and the matching method based on feature points can generally obtain stable matching feature points, we compared the proposed method with four state-of-the-art feature matching methods, including SIFT [44], Harris [45], local linear transform (LLT) [46], and fast dense feature matching (FD) [47]. For our pre-matching method, we normalized the pre-matching cost to [0,1], eliminating matching points whose cost was greater than the threshold value of 0.01 to retain the matching points with high confidence. The results of the comparison experiment using Dataset A are shown in Figure 8. The red points in the image show the extracted feature points. The number and accuracy of matching points obtained by the five methods are compared, as shown in Table 1, where the points represent the average number of matching points extracted from each set of images.

Figure 8. Some pre-matching results with different methods. (**a**) Results of SIFT; (**b**) results of Harris; (**c**) results of LLT; (**d**) results of FD; (**e**) results of pre-matching method. The top and bottom are right and left views, respectively.

Table 1. Comparison of pre-matching results with different methods.

Method	EPE	RMSE	D1	D3	Points
SIFT	1.31	3.15	0.17	0.10	984
Harris	1.12	2.44	0.19	0.06	319
LLT	1.24	2.75	0.17	0.07	973
FD	1.35	3.45	0.17	0.11	1115
Proposed	1.11	2.45	0.12	0.06	29,021

The comparison results of Figure 8 and Table 1 clearly show that the number of matching points of SIFT, Harris, LLT and FD is small and mainly distributed on the edges of buildings and roads with rich textures; there are few matching points on the roofs of buildings and road surfaces with weaker textures. In contrast, our proposed pre-matching method is able to provide more dense matching points on buildings and roads that are of more concern in urban areas. This is due to the fact that the matching cost of our proposed

pre-matching method is able to achieve aggregation by surrounding superpixel blocks, thus ensuring that weakly textured areas also provide stable matching points. Our proposed pre-matching method extracts the largest number of matched point pairs, and the accuracy is close to that of other feature matching methods, which can provide sufficient training samples for the whole self-supervised matching.

4.3. Ablation Study of SRWP and PCAM

Considering that our proposed matching method contains several key links, and the adjustment of each link will have an impact on the performance of SRWP and PCAM, we compared the effects of different structures or parameters on the method through ablation experiments.

- Pre-matching results

To verify the impact of different confidence levels and number of labeled samples obtained by pre-matching on the overall algorithm, we set the pre-matching cost threshold to [0.001,0.005,0.01,0.02,0.05] and conducted the experiments separately. The experiments were conducted using Dataset A, and the result is shown in Table 2. The number of points is the number of labeled training samples provided for pre-matching, and EPE/D1/D3 metrics are the average of 1685 sets of experimental data.

Table 2. Comparison of different pre-matching results.

Method	Threshold	Number of Points	EPE	D1	D3
Pre-matching	0.001	20,922	2.53	0.28	0.18
Pre-matching	0.005	24,576	2.50	0.27	0.18
Pre-matching	0.01	29,021	2.44	0.25	0.16
Pre-matching	0.02	39,097	2.62	0.29	0.19
Pre-matching	0.05	51,724	2.79	0.32	0.23

The matching accuracy comparison in Table 2 shows that the setting of the pre-matching threshold will affect the performance of SRWP and PCAM. As the threshold increases, the more labeled samples are extracted, but the quality of these samples decreases continuously. Therefore, the threshold is not set as high as possible or as low as possible. In this paper, we set the threshold to 0.01 through several experiments.

- Different Loss

We evaluated the method using different losses to test the effectiveness of the loss function. The experiments were conducted using Dataset A. The EPE/D1/D3 metrics are the average of 1685 sets of experimental data. The experimental results are shown in Table 3.

Table 3. Comparison results with different losses.

L_p	L_s	L_{PCAM}	L_l	EPE	D1	D3
✓				5.79	0.31	0.24
✓	✓			4.12	0.30	0.22
✓	✓	✓		3.24	0.28	0.19
✓	✓	✓	✓	2.44	0.25	0.16

Four kinds of losses are defined in this paper. Table 3 shows that the evaluation metrics EPE/D1/D3 are relatively high if only photometric loss is used for training. This is because of the structural similarity of the photometric loss description; the weak texture areas cannot be well handled. If the loss includes smoothness loss, the evaluation metrics achieve a certain decrease, which is due to the fact that non-disparity abrupt regions in remote sensing images still occupy the majority of the image. If PCAM loss is added, the

performance will gradually improve, which is due to the fact that the loss fully takes into account the left–right consistency and cyclic consistency existing in stereo images. Finally, adding L_1 loss as labeled samples using matching points with high confidence obtained by pre-matching results in a significant decrease in the evaluation metrics.

4.4. Flexibility of SRWP and PCAM

Considering that SRWP and PCAM will encounter various conditions of data in practical application, we group the experimental data to compare the effectiveness of the method in dealing with different data.

- Resolutions

To test the flexibility of the method to different image resolutions, we adjusted both the test image and the true disparity map of Dataset A to 1024 × 1024, 512 × 512, 256 × 256, and 128 × 128 resolutions, and the values in the true disparity map are adjusted to the corresponding scales. The results are shown in Table 4.

Table 4. Comparison results with different resolutions.

Resolutions	EPE	D1	D3
1024 × 1024	2.44	0.25	0.16
512 × 512	2.35	0.23	0.15
256 × 256	2.36	0.22	0.14
128 × 128	2.34	0.20	0.14

Table 4 shows that the performance of our method has little change under the four resolutions, because our method calculates the disparity for pixels, which are less disturbed by the resolution. On the other hand, as the resolution becomes smaller, there is a certain improvement in matching accuracy, which is due to the fact that the range of disparity is continuously reduced during down-sampling, thus reducing the matching difficulty.

- Maximum Parallax

In order to test the flexibility of the method for different disparity sizes, we grouped the test images of Dataset A into four groups for testing separately, and the upper and lower bounds of disparity for the images in each group were (0,40), (0,80), (0,120), and (0,160). The results obtained are shown in Table 5.

Table 5. Comparison results with different maximum disparity.

Range of Disparity	EPE	D1	D3
(0,40)	2.42	0.24	0.16
(0,80)	2.56	0.26	0.18
(0,120)	2.64	0.27	0.20
(0,160)	2.80	0.28	0.21

As can be observed from Table 5, the performance of our method does not vary much for different disparity ranges, which is due to the fact that our method is able to handle large disparity differences by means of a saliency map. However, the overall trend is that the larger the saliency map range, the worse the accuracy, which is due to the fact that the larger the saliency map range, the more difficult it is to search for the true saliency map value. In general, our method can ensure that the matching accuracy decreases slightly while the disparity range increases significantly, which proves that our method can adapt to different parallax ranges.

4.5. Results and Discussion

In this section, experiments are conducted using Datasets A, B, and C, where Datasets A and B utilize true disparity values and evaluate disparity accuracy by EPE, D1, and D3 metrics. Since there is no corresponding disparity true value for Dataset C, we use the matching results to 3D reconstruction. We calculate the RMSE of DSM obtained by three-dimensional reconstruction, and indirectly evaluate the matching accuracy. In order to fully demonstrate the effectiveness of the proposed algorithm, some of the latest algorithms are selected for comparison experiments, including three types of traditional methods (FCVFSM [13], SGM [10] and SGBM [24]) and three types of neural network methods (PSMNet [19], CGN [48] and BGNet [49]). The parameters of all compared methods are set according to the recommendations of their articles, the network models are used in their original models, and the results of the quantitative evaluation are shown in Table 6, where time refers to the test time.

Table 6. Comparison of different stereo matching methods.

Method	Dataset A				Dataset B			Dataset C	
	EPE	D1	D3	Time (s)	EPE	D1	D3	RMSE	Time (s)
FCVFSM	4.56	0.49	0.36	24.67	4.84	0.47	0.39	6.21	3.79
SGBM	4.95	0.47	0.32	22.96	5.10	0.45	0.35	7.19	3.53
SGM	3.73	0.40	0.29	8.95	3.65	0.39	0.27	5.14	1.37
PSMNet	3.14	0.33	0.26	1.36	3.23	0.34	0.25	3.75	0.21
CGN	3.39	0.35	0.22	1.12	3.45	0.38	0.24	3.93	0.17
BGNet	2.85	0.31	0.18	1.10	2.83	0.31	0.17	3.32	0.17
Proposed	2.44	0.25	0.16	1.18	2.32	0.24	0.14	2.36	0.18

The first set of experiments is from Dataset A, containing 1685 pairs of stereo images, each 1024 × 1024 pixels. Some experimental results are shown in Figure 9.

It can be seen from the experimental results in Figure 9 that our method is visually closer to the true value of disparity than other methods. At the same time, our method is able to suppress the singular value generation. This is due to the smoothing loss set by our method and the fact that training at different resolutions can effectively weight the singular errors.

The second set of experiments is from Dataset B. Stereo matching experiments are performed separately, and then the matched disparity maps are multiplied by the mask to obtain disparity map results for weak texture or parallax abrupt change regions; some of the experimental results are shown in Figure 10.

The experimental results in Figure 10 show that our method is closer to the true value of disparity compared with other methods in the region of weak texture or parallax abrupt change. The green boxes marking the edges of the buildings show that our method can accurately locate the disparity faults between the buildings and the ground, ensuring the disparity edges are flush and closer to the true disparity value in terms of shape contours. This is because our method not only ensures that the training samples contain a large number of disparity mutation edges through pre-matching, but also optimizes the disparity edges through disparity refinement. As can be seen from the top of the building, marked by the black circle, our method is able to obtain satisfactory disparity results in the weakly textured regions, maintaining disparity smoothness in the flat areas on top of the building. This is mainly due to the loss of smoothness and PCAM, which limits the generation of distortion. The experimental results in Figure 10 demonstrate that our method can handle stereo matching of weak textures and parallax abrupt change regions.

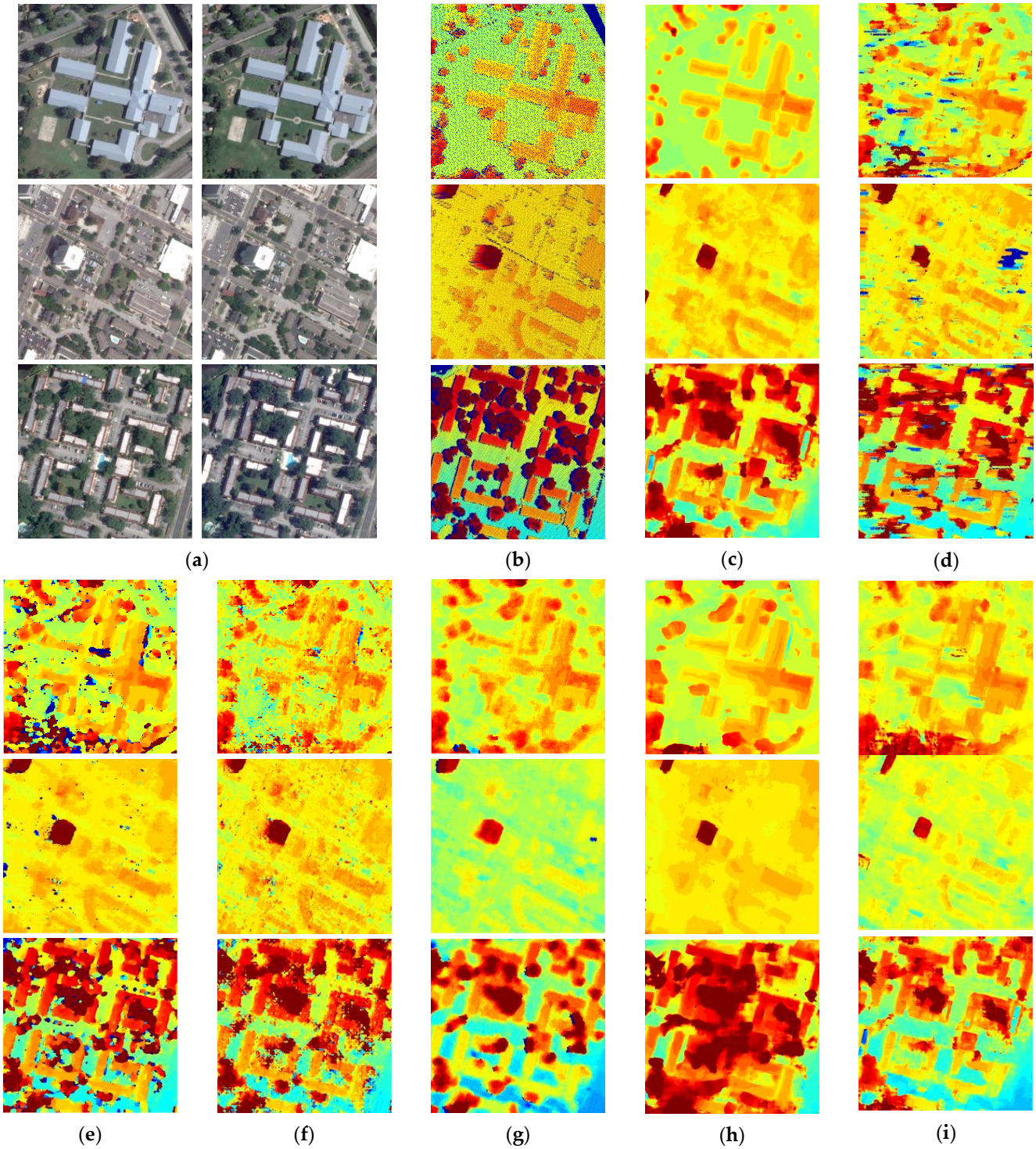


Figure 9. Some stereo matching results with different methods for Dataset A. (**a**) Optical satellite stereo pairs; (**b**) truth of disparity; (**c**) results of proposed method; (**d**) results of FCVFSM; (**e**) results of SGBM; (**f**) results of SGM; (**g**) results of CGN; (**h**) results of BGNet; (**i**) results of PSM.

The third set of experiments from Dataset C contains a set of stereo images of the Harbin area with image pixel sizes 21,691 × 15,069 and 22,271 × 15,172. The results of the partially reconstructed DSM experiments are shown in Figure 11. The size of this area is 400 × 400 pixels. It can be seen from the figure that our method is able to compute accurate disparity maps of satellite images and reliably reconstruct the DSMs based on our disparity results.

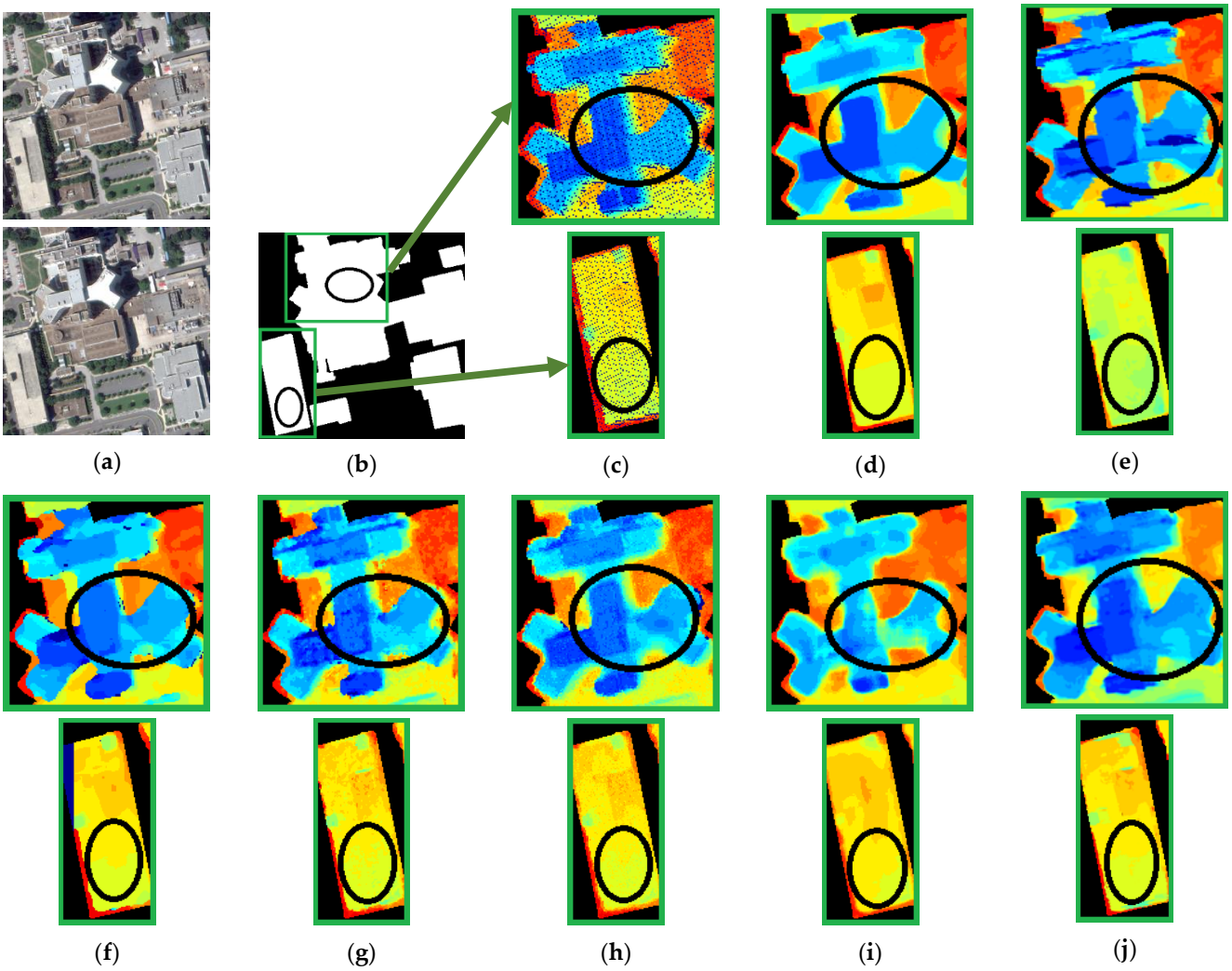


Figure 10. Some stereo matching results with different methods for dataset B. (**a**) Optical satellite stereo pairs; (**b**) mask of left image; (**c**) truth of disparity; (**d**) results of proposed algorithm; (**e**) results of FCVFSM; (**f**) results of SGBM; (**g**) results of SGM; (**h**) results of CGN; (**i**) results of BGNet; (**j**) results of PSM.

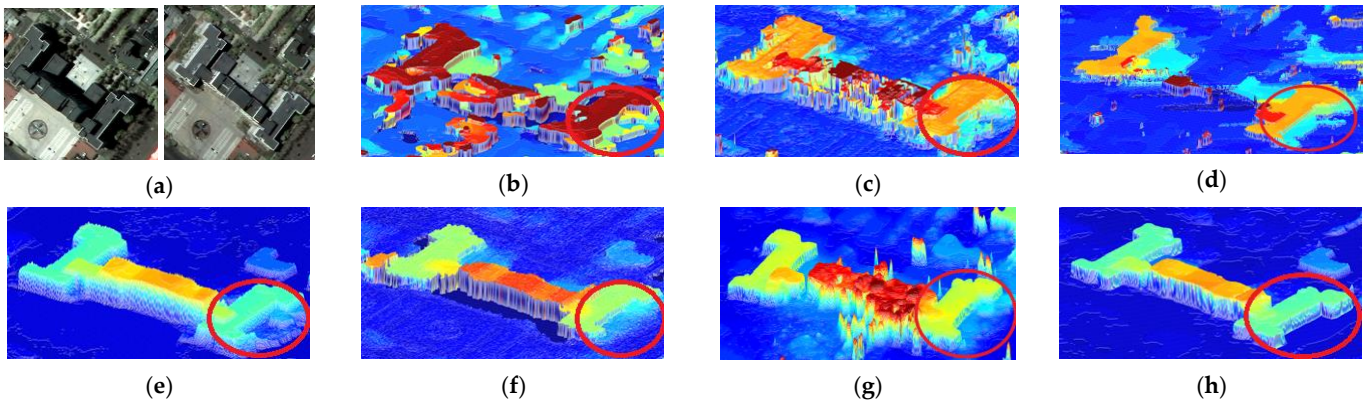


Figure 11. Some 3D reconstruction results with different methods for dataset C. (**a**) Optical satellite stereo pairs; (**b**) results of SGBM; (**c**) results of SGM; (**d**) results of FCVFSM; (**e**) results of CGN; (**f**) results of BGNet; (**g**) results of PSM; (**h**) results of proposed algorithm.

The experimental results in Figure 11 show that our method can maintain the flatness of the building roof compared with other methods, and also the reconstructed building has a more regular shape. The edges of the building marked by the red circles show that our method can achieve the separation of the building from the ground and effectively avoid the error caused by the difference in viewpoint.

We use Dataset A, B and C for experimental analysis and compare with state-of-the-art. From the experimental results in Figures 9–11 and Table 6, it can be seen that SGM as a classic dynamic programming stereo matching algorithm can only obtain a rough disparity map when it involves parallax abrupt change and building occlusion. This is due to the difficulty of dynamic programming methods to compute edge pixel disparity only through the optimization of cost aggregation. As a variant of SGM, SGBM can improve the smoothness of parallax edges but cannot handle large areas without texture. As shown in Figure 10f,g, SGBM performs worse in untextured regions marked by black circles. FCVFSM assumes that the pixels in the support window have a constant parallax value, which violates the inclined surface and causes the edge to blur in the parallax abrupt change region. CGN achieves unsupervised matching by means of generative adversarial networks. However, its generator does not consider the attention mechanism and cannot improve the matching accuracy of buildings in the urban areas we are concerned about. BGnet and PSMNet focuses on the improvement of the network structure and still needs a large number of training samples containing ground-truth labels. This experiment directly uses its original network model, and it is difficult to accurately find the matching points of trees, buildings and other targets on the remote sensing images.

In terms of time, traditional methods take longer than neural network methods. The method proposed in this paper has no advantage in test time. However, considering that this paper is oriented to 3D reconstruction under mass satellite remote sensing data, in order to reduce the cost of manually labeling samples; it does not need to be as fast as the stereo matching driverless method.

Our method is closer to the true value of disparity than other methods, whether in the whole image, weak texture region or parallax abrupt change region. It also has more advantages in accuracy comparison. The metrics value of the proposed method is the lowest among the five algorithms. Since all the test data are satellite remote sensing images of urban areas, the proposed self-supervised stereo matching method based on SRWP and PCAM is more suitable for urban satellite remote sensing images than other methods.

5. Conclusions

Stereo matching and 3D reconstruction of urban scenes using optical satellite remote sensing stereo images are more advantageous in terms of data cost and coverage range. However, the urban scenes are complex and diverse, and there are problems such as occlusion of building entities caused by different viewpoints, weak textures and parallax abrupt change. Traditional matching methods using only simple features have difficulty achieving better results. In contrast, the stereo matching method based on neural networks needs to use a large number of training samples containing true labels, which are difficult to obtain from satellite images. To this end, this paper proposes a self-supervised stereo matching method based on SRWP and PCAM to achieve stereo matching of urban scenes. First, based on the idea of matching cost optimization and updating, a superpixel random walk pre-matching method is proposed. The initial point matching cost is constructed by selecting simple features, and aggregated into a block matching cost based on the superpixel segmentation results. The matching cost is optimized and updated, and the pre-matching point pairs with high confidence are filtered using thresholding. We then introduce parallax attention and channel attention into stereo matching and construct a parallax-channel attention mechanism to capture the correspondence of stereo images. We also add two optimization strategies, feature enhancement and parallax refinement, to obtain more significant features in pre-processing and more refined parallax in post-processing, respectively. The network model is then trained using pre-matching point pairs

and various unsupervised loss constraints. Finally, the trained network is used to re-match the stereo images to obtain disparity map and stereo matching relationships. Comparative experiments and analyses on publicly and practically measured constructed datasets show that our method is able to learn the parallax correspondence in a self-supervised manner and can achieve more advanced performance compared to other methods. However, there is still some interference error for vegetation such as trees. Future research will focus on eliminating the effect of tree interference on stereo matching and 3D reconstruction.

Author Contributions: Methodology, W.C.; project administration, H.C.; data curation, S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 61771170.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Xiao, X.; Guo, B.; Li, D. Multi-view stereo matching based on self-adaptive patch and image grouping for multiple unmanned aerial vehicle imagery. *Remote Sens.* **2016**, *8*, 89. [CrossRef]
2. Nguatem, W.; Mayer, H. Modeling urban scenes from Pointclouds. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3857–3866.
3. Wohlfeil, J.; Hirschmüller, H.; Piltz, B.; Börner, A.; Suppa, M. Fully automated generation of accurate digital surface models with sub-meter resolution from satellite imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *XXXIX-B3*, 75–80. [CrossRef]
4. Zhao, L.; Liu, Y.; Men, C.; Men, Y. Double propagation stereo matching for urban 3-D reconstruction from satellite imagery. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–17. [CrossRef]
5. Zhou, C.; Zhang, H.; Shen, X. Unsupervised learning of stereo matching. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1567–1575.
6. Pang, J.; Sun, W.; Yang, C. Zoom and learn: Generalizing deep stereo matching to novel domains. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2070–2079.
7. Muresan, M.P.; Nedevschi, S.; Danescu, R. A multi patch warping approach for improved stereo block matching. In Proceedings of the International Conference on Computer Vision Theory and Applications, Porto, Portugal, 27 February–1 March 2017; pp. 459–466.
8. Spangenberg, R.; Langner, T.; Adfeldt, S.; Rojas, R. Large scale semi-global matching on the CPU. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV 2014), Dearborn, MI, USA, 8–11 June 2014; pp. 195–201.
9. Liu, X.; Li, Z.H.; Li, D.M. Computing stereo correspondence based on motion detection and graph cuts. In Proceedings of the 2012 Second International Conference on Instrumentation, Measurement, Computer, Communication and Control (IMCCC), Harbin, China, 8–10 December 2012; pp. 1468–1471.
10. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. [CrossRef]
11. Huang, X.; Zhang, Y.; Yue, Z. Image-guided non-local dense matching with three-steps optimization. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *3*, 67–74. [CrossRef]
12. Li, M.; Kwoh, L.K.; Yang, C.-J.; Liew, S.C. 3D building extraction with semi-global matching from stereo pair worldview-2 satellite imageries. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, 26–31 July 2015; pp. 3006–3009. [CrossRef]
13. Rhemann, C.; Hosni, A.; Bleyer, M.; Rother, C.; Gelautz, M. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 504–511.
14. Oh, C.; Ham, B.; Sohn, K. Probabilistic correspondence matching using random walk with restart. In Proceedings of the British Machine Vision Conference (BMVC 2012), Guildford, UK, 3–7 September 2012; pp. 1–10.
15. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
16. Shaked, A.; Wolf, L. Improved stereo matching with constant highway networks and reflective confidence learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6901–6910. [CrossRef]
17. Zbontar, J.; LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1592–1599.

18. Kendall, A.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; Bry, A. End-to-end learning of geometry and context for deep stereo regression. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 66–75.
19. Chang, J.-R.; Chen, Y.-S. Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 5410–5418.
20. Seki, A.; Pollefeys, M. SGM-Nets: Semi-global matching with neural networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 6640–6649.
21. Suliman, A.; Zhang, Y. Double projection planes method for generating enriched disparity maps from multi-view stereo satellite images. *Photogramm. Eng. Remote Sens.* **2017**, *83*, 749–760. [CrossRef]
22. Tatar, N.; Saadatseresht, M.; Arefi, H.; Hadavand, A. Quasi-epipolar resampling of high resolution satellite stereo imagery for semi global matching. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *40*, 707. [CrossRef]
23. Mandanici, E.; Girelli, V.A.; Poluzzi, L. Metric accuracy of digital elevation models from worldview-3 stereo-pairs in urban areas. *Remote Sens.* **2019**, *11*, 878. [CrossRef]
24. Yang, W.; Li, X.; Yang, B.; Fu, Y. A novel stereo matching algorithm for digital surface model (DSM) generation in water areas. *Remote Sens.* **2020**, *12*, 870. [CrossRef]
25. Zhu, H.; Jiao, L.; Ma, W.; Liu, F.; Zhao, W. A novel neural network for remote sensing image matching. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2853–2865. [CrossRef] [PubMed]
26. Tao, R.; Xiang, Y.; You, H. Stereo matching of VHR remote sensing images via bidirectional pyramid network. In Proceedings of the IGARSS 2020–2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa Village, HI, USA, 16–26 July 2020; pp. 6742–6745.
27. Froba, B.; Ernst, A. Face detection with the modified census transform. In Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, 19 May 2004; pp. 91–96.
28. Norouzi, M.; Fleet, D.J.; Salakhutdinov, R. Hamming distance metric learning. *Adv. Neural Inf. Process. Syst.* **2012**, *2*, 1061–1069.
29. Yin, J.; Wang, T.; Du, Y.; Liu, X.; Zhou, L.; Yang, J. SLIC superpixel segmentation for polarimetric SAR images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–17. [CrossRef]
30. Dong, X.; Shen, J.; Shao, L.; Van Gool, L. Sub-markov random walk for image segmentation. *IEEE Trans. Image Process.* **2016**, *25*, 516–527. [CrossRef]
31. Li, W.; Qi, F.; Tang, M.; Yu, Z. Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing* **2020**, *387*, 63–77. [CrossRef]
32. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
33. Wang, L.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W.; Guo, Y. Learning parallax attention for stereo image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12250–12259.
34. Li, H.; Qiu, K.; Chen, L.; Mei, X.; Hong, L.; Tao, C. SCAttNet: Semantic segmentation network with spatial and channel attention mechanism for high-resolution remote sensing images. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 905–909. [CrossRef]
35. Zhang, K.; Fang, Y.; Min, D.; Sun, L.; Yang, S.; Yan, S.; Tian, Q. Cross-scale cost aggregation for stereo matching. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1590–1597.
36. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2018; pp. 85–100.
37. Guo, C.; Chen, D.; Huang, Z. Learning efficient stereo matching network with depth discontinuity aware super-resolution. *IEEE Access* **2019**, *7*, 159712–159723. [CrossRef]
38. Fleet, D.; Weiss, Y. Optical flow estimation. In *Handbook of Mathematical Models in Computer Vision*; Springer: Boston, MA, USA, 2006; pp. 239–257.
39. Godard, C.; Mac Aodha, O.; Brostow, G.J. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 270–279.
40. Li, A.; Yuan, Z. Occlusion aware stereo matching via cooperative unsupervised learning. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 4–6 December 2018; pp. 197–213.
41. Yin, Z.; Shi, J. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1983–1992.
42. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
43. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:abs/1412.6980.
44. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
45. Ram, P.; Padmavathi, S. Analysis of Harris corner detection for color images. In Proceedings of the 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, India, 3–5 October 2016; pp. 405–410.
46. Ma, J.; Zhou, H.; Zhao, J.; Gao, Y.; Jiang, J.; Tian, J. Robust feature matching for remote sensing image registration via locally linear transforming. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 6469–6481. [CrossRef]

47. Du, W.-L.; Li, X.-Y.; Ye, B.; Tian, X.-L. A fast dense feature-matching model for cross-track pushbroom satellite imagery. *Sensors* **2018**, *18*, 4182. [CrossRef] [PubMed]
48. Pilzer, A.; Xu, D.; Puscas, M.; Ricci, E.; Sebe, N. Unsupervised adversarial depth estimation using cycled generative networks. In Proceedings of the 2018 International Conference on 3D Vision (3DV), Verona, Italy, 5–8 September 2018; pp. 587–595.
49. Xu, B.; Xu, Y.; Yang, X.; Jia, W.; Guo, Y. Bilateral grid learning for stereo matching network. *arXiv* **2021**, arXiv:abs/2101.01601.