remote sensing

MDPI

Technical Note

UAV Remote Sensing Image Automatic Registration Based on Deep Residual Features

Xin Luo [1,2], Guangling Lai [1,2], Xiao Wang [1], Yuwei Jin [1,2], Xixu He [1], Wenbo Xu [1,2] and Weimin Hou [3,*]

[1] School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 611731, China; luoxin@uestc.edu.cn (X.L.); 202022070206@std.uestc.edu.cn (G.L.); wangxiao29@std.uestc.edu.cn (X.W.); 201711180104@std.uestc.edu.cn (Y.J.); HL@uestc.edu.cn (X.H.); xuwenbo@uestc.edu.cn (W.X.)

[2] Yangtze Delta Region Institute (Huzhou), University of Electronic Science and Technology of China, Huzhou 313001, China

[3] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China

* Correspondence: hwm@hebust.edu.cn; Tel.: +86-133-6386-1805

Abstract: With the rapid development of unmanned aerial vehicle (UAV) technology, UAV remote sensing images are increasing sharply. However, due to the limitation of the perspective of UAV remote sensing, the UAV images obtained from different viewpoints of a same scene need to be stitched together for further applications. Therefore, an automatic registration method of UAV remote sensing images based on deep residual features is proposed in this work. It needs no additional training and does not depend on image features, such as points, lines and shapes, or on specific image contents. This registration framework is built as follows: Aimed at the problem that most of traditional registration methods only use low-level features for registration, we adopted deep residual neural network features extracted by an excellent deep neural network, ResNet-50. Then, a tensor product was employed to construct feature description vectors through exacted high-level abstract features. At last, the progressive consistency algorithm (PROSAC) was exploited to remove false matches and fit a geometric transform model so as to enhance registration accuracy. The experimental results for different typical scene images with different resolutions acquired by different UAV image sensors indicate that the improved algorithm can achieve higher registration accuracy than a state-of-the-art deep learning registration algorithm and other popular registration algorithms.

Keywords: UAV image; registration; ResNet; deep residual feature; PROSAC

Citation: Luo, X.; Lai, G.; Wang, X.; Jin, Y.; He, X.; Xu, W.; Hou, W. UAV Remote Sensing Image Automatic Registration Based on Deep Residual Features. *Remote Sens.* **2021**, *13*, 3605. https://doi.org/10.3390/rs13183605

Academic Editor: Liang-Jian Deng

Received: 29 July 2021
Accepted: 6 September 2021
Published: 10 September 2021

1. Introduction

Nowadays, unmanned aerial vehicles (UAVs) are often used to collect airborne remote sensing images. However, the view fields of drone images are often limited by flight heights and camera focal lengths. As a result, it is commonly impossible to display an entire study area through a single image. In this case, image registration technology can assemble several single images with overlapping areas according to their own feature information to yield a large-scope scene image for subsequent scientific researches or applications [1–3]. Hence, UAV image registration is widely used in scene mosaicking and panorama production, so as to integrate information of acquired images and make up for the shortcomings of UAV photography.

In the field of image registration, algorithms based on feature points are most popular. In 1999, the scale invariant feature transform (SIFT) operator was proposed by D.G. Lowe et al. [4]. Features that are invariant to image scale, rotation and scaling can be obtained by using this method, so it has been widely used [5–7]. Since it takes a long time for the SIFT algorithm to yield feature descriptors of 128 dimension, some scholars have proposed different improved versions. The most famous is the speed up robust features

(SURF) algorithm, which was proposed by H. Bay et al. in 2006 [8]. It utilizes a wavelet transform to construct feature vectors and reduces the dimension of vectors to 64, which improves registration speeds. At present, the SURF algorithm has also become one of the most commonly used registration algorithms. It has been widely adopted in many fields, such as fingerprint registration [9], vehicle monitoring [10], and precision agriculture [11]. Furthermore, the KAZE algorithm uses a nonlinear scale space instead of the Gaussian difference scale space to detect feature points and construct feature description vector. The features extracted by this method have good adaptability to condition changes, such as illumination and rotation [12]. Lately, a fast local feature detection operator, the Oriented FAST and rotated BRIEF (ORB), is proposed by Ethan Rublee et al. in 2011. It can provide fast calculation speeds, strong real-time performances and robustness to image noise. However, it is inferior to the KAZE algorithm and the SIFT algorithm in terms of accuracy.

A deep convolutional neural network (CNN) proposed in 1998 that was developed to solve numeral recognition problems [13]. With recent rapid development, deep neural networks have appeared in image classification, semantic segmentation, target detection and other fields. They possess promising application prospects. The methods based on deep neural networks also exhibit excellent performances in image registration. For instance, convolutional neural networks are used to regress homography parameters for UAV multispectral image registration, and pyramid structure similarity loss is used to optimize the networks [14]. Unlabeled data were utilized to train a convolutional neural network and extract image features. Compared with traditional feature-based image registration algorithms, this method constructed a feature description vector that resulted in a better performance [15]. In addition, the ORB algorithm is applied to extract local features for rough localization, and then shapes obtained by the U-Net semantic segmentation are employed in fine matching [16]. Moreover, a convolutional neural network directly was used to learn transform parameters for image registration in order to simplify the process of image registration [17]. In some research studies, SIFT features and CNN features were fused in order to provide a large amount of middle-level and high-level information for remote-sensing image registration. This scheme has better efficiency and can gain more correct matched point pairs to improve registration accuracy [18]. Some researchers use images to be registered to train a neural network and effectively segment the incomplete/defective brain data. On this basis, they register different temporal rat brain images [19]. The abovementioned methods are still completely or partially dependent on the visual contents of images. Some of them need extra training and can only process images of specific objects. It is rarely reported that research studies on automatic registration directly use high-level abstract features from deep networks instead of low-level features of images.

As for UAV remote-sensing images, owing to their high resolution and complex details, many mistakes will occur in detecting and matching image feature points. Therefore, research on registration methods for UAV remote-sensing image is challenging and valuable. How to register UAV images with high accuracy has become a hot topic in image processing. Structural consistency between the images to be matched are exploited in a registration strategy considering intensive illumination and contrast changes in multi-temporal UAV images [20]. This method is verified independently of radiation information and is efficient in multi-temporal UAV image registration. A multi-view image registration method for small drones was utilized to extract change information of arable land in hilly areas, and it achieved the expected results [21].

This work presents an UAV image automatic registration method based on ResNet-50 [22], a popular deep residual neural network. This method does not rely on low-level image features, such as points, lines and shapes, and needs no extra training to achieve high registration accuracy. ResNet-50 serves as a feature-extraction network. The feature aggregation by Kronecker product are adopted in feature vector construction. The progressive sampling consensus (PROSAC) algorithm is utilized to remove false matches and fit registration parameters. In experiments, the proposed method is compared with

another prominent registration method based on deep learning and current representative registration methods based on point features. The rest of this article is organized as follows. Section 2 introduces our UAV image registration strategy in detail. The experimental results on images of different scenes and from different sources are given and analyzed in Section 3. Section 4 is the conclusion of this article.

2. Materials and Methods

Figure 1 illustrates the registration pipeline of our proposed method based on deep residual features. First of all, the deep residual network neural network ResNet-50 is exploited to extract feature for image registration. Here, the center of each 8 × 8 pixel region of an image is taken as a feature point. A multi-scale feature description vector of the feature points is constructed through convolution-layer outputs of the feature-extraction network in the ResNet-50 architecture. The outputs of residual blocks are merged by the Kronecker product to construct multi-scale feature description vectors. After feature matching, the PROSAC algorithm is utilized to remove false mismatches and fit a geometric transform model.

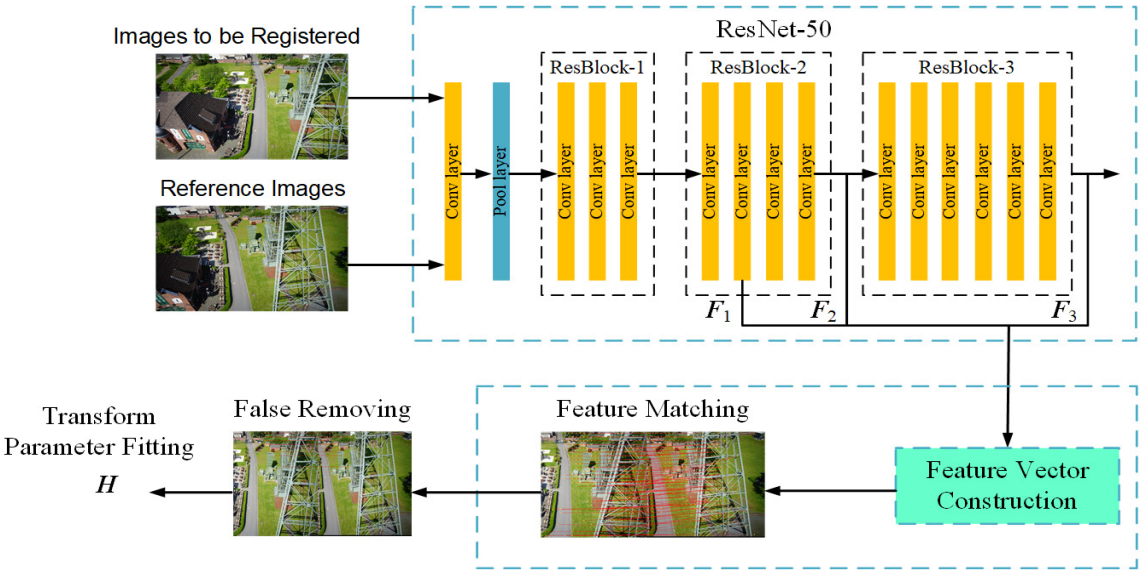


Figure 1. The registration pipeline of our proposed method based on deep residual features. Firstly, three feature map $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, and $\boldsymbol{F}_3$ corresponding to an 8 × 8 pixel region in UAV images are extracted from ResNet-50. Then they are feature integrated by Kronecker product to construct feature vectors. Finally, the PROSAC algorithm is performed to remove false matches and yield a homography matrix, $\boldsymbol{H}$, for registration transform.

In this work, inspired by a registration method based on a deep convolutional neural network VGG-16 [23], feature vectors of a deep residual network are built according to characteristics of UAV remote sensing images. Hence, high-level abstract features of images play main roles in registration processes. In VGG-16 feature construction, the lower-level image features outputted by the pooling layer 1 (pool1) and the pooling layer 2 (pool2) are discarded. The feature description vectors are assembled on the basis of the output features of three high-level layers, i.e., the pooling layer 3 (pool3), the pooling layer 4 (pool4) and a self-defined pooling layer (pool5-1). Therefore, similarly, we design registration feature vectors of ResNet-50 through a hidden-layer output of ResBlock-2, and the output of ResBlock-2 and ResBlock-3. Our strategy can overcome the degradation problem that appears with deepening neural networks, so that extracted image features are more representative.

2.1. Deep Residual Feature Extraction Network

ResNet emerged in 2015. By virtue of its depth and simple structures, this network has become more and more popular. It has shown stronger classification and detection capacity than the VGG series networks. Theoretically, as a network becomes deeper, the extracted image features will be more representative. However, blindly deepening the network will result in a learning efficiency decline of networks, and accuracy of tasks will cease to increase, or even decrease. The development of ResNet is primarily to cope with the degradation problem as networks are deepened [22]. As a residual unit depicted in Figure 2, it can be found that there is an extra curve, i.e., a skip connection, from the input to the output. By this means, the ResNet networks can learn differences between inputs to outputs. That is, they learn residual changes instead of fitting functions. Even with continuously increasing of network depth, the ResNet networks still have good sensitivities to residuals, which can avoid the problem of gradient vanishing or explosion. This design makes the training processes of networks sample and fast. Hence, the invention of residual structures can greatly increase the depth of deep learning networks without over-fitting [24].

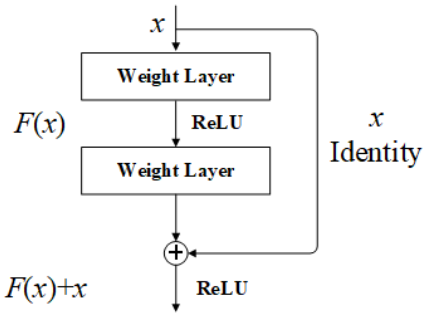


Figure 2. A residual block structure [22].

A residual unit can be expressed by the following:

$$y_l = h(x_l) + F(x_l, W_l) \tag{1}$$

and

$$x_{l+1} = f(y_l) \tag{2}$$

where x_l and x_{l+1} respectively represent the input and output eigenmatrix of the lth residual element; F is a residual function, representing the learned residual via the unit; $h(x_l) = x_l$ represents an identity mapping; and f is an ReLU activation function. Based on the above formulas, the feature learned from a shallower layer (l) to a deeper layer (L) is given by the following:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{3}$$

According to chain rules, the backpropagation gradient can be written as follows:

$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot [1 + \frac{\partial}{\partial x_L} F(x_i, W_i)] \tag{4}$$

where $\partial loss/\partial x_l$ represents the gradient of a loss function before the *l-th* layer, and the 1 in the bracket represents the lossless transfer of the gradient by the shortcut mechanism of residual networks. The other residual gradient needs to go through the weights layer and cannot be transferred directly. As a result, the residual gradients will not all be -1. It is most important that when their values become smaller as the accumulation of networks, the existence of 1 can restrain vanishing gradients. Therefore, compared with the ordinary deep learning architecture, it is easier to learn from residuals. In summary, introducing

shortcuts enables identity mappings to be realized in ResNet, and, in this way, gradients can be transferred smoothly among different layers.

The architectures of various ResNet networks are basically similar. Data pass through a $7 \times 7 \times 64$ convolution layer, a 3×3 max-pooling layer for down-sampling, various residual layers and an average pooling layer for down-sampling, successively. In the end, a Softmax model converts the previous outputs into a probability distribution to yield the final output. In general, the depth of ResNet networks are 18, 34, 50, 101 and 152. Among them, ResNet-50 and ResNet-101 are more common. The calculation complexity and time cost of the ResNet-101 network are relatively high. Hence, the other representative ResNet-50 with moderate number of layers are chosen for feature extraction in image registration. As illustrated in Figure 3, the feature-extraction network of ResNet-50 consists of five stages. Each stage is composed of a different number of residual blocks, and each residual block is realized via three convolution layers to eliminate depth effects. The number of learnable parameters in the ResNet-50 network model is up to 23 million.

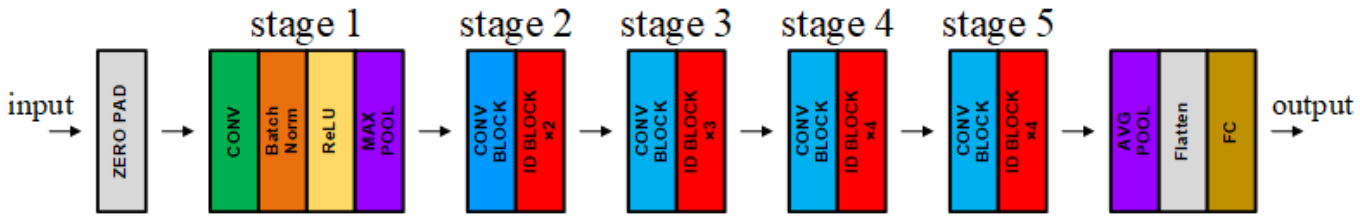


Figure 3. The ResNet-50 Block Architecture [20].

2.2. Feature Description Vector Construction and Matching

2.2.1. Feature Description Vector Construction

At first, since UAV images are usually of high resolution, if they are down-sampled at the input side of a deep learning network, image feature information will mostly be lost and, consequently, registration errors will increase. Therefore, an arbitrary-size image to be registered and its reference image of the same size are inputted into the ResNet-50 feature-extraction network with their original resolutions. Each region of 8×8 pixels in an input image is defined as a feature point. Accordingly, the 8×8, 16×16 and 32×32 pixel regions around a feature point are employed to extract feature vectors at different scales, which corresponds to the output of ResBlock-1 (residual block 1), ResBlock-2 (residual block 2) and ResBlock-3 (residual block 3) in ResNet-50, respectively. However, the features extracted by ResBlock-1 belong to the low level. Thus, a middle-layer output of ResBlock-2 was adopted in this work to construct a feature description vector, together with the output of ResBlock-2 and of ResBlock-3, for registration.

By some comparison, it is revealed that using the output of the second convolutional layer in ResBlock-2 as the first feature map, $\boldsymbol{F}_1$, to construct a feature description vector for a feature point, corresponding to an 8×8 pixel region in input images, can achieve an ideal registration performance. Given that the size of an input image is $N \times N$, the size of $\boldsymbol{F}_1$ is $(N/8) \times (N/8) \times 512$. Every 8×8 pixel region in the input image corresponds to a 512-dimensional vector in $\boldsymbol{F}_1$. On the other hand, each 16×16 pixel region in the input image corresponds to a 512-dimensional vector in the output of ResBlock-2, denoted by $\boldsymbol{O}_{\text{ResBlock-2}}$. Hence, the size of $\boldsymbol{O}_{\text{ResBlock-2}}$ is $(N/16) \times (N/16) \times 512$ pixels. Since one feature vector in ResBlock-2 is shared by four defined feature points, the Kronecker product, denoted by a $\otimes$ symbol, is performed on $\boldsymbol{O}_{\text{ResBlock-2}}$ to obtain the second feature map, $\boldsymbol{F}_2$, for an input image:

$$\boldsymbol{F}_2 = \boldsymbol{O}_{\text{ResBlock}-2} \otimes \boldsymbol{I}_{2\times 2\times 1} \tag{5}$$

where $\boldsymbol{I}$ represents a tensor of subscripted shapes and it is filled with unities. Given that $\boldsymbol{A} = (a_{ij}) \in c^{m \times n}$ and $\boldsymbol{B} = (b_{ij}) \in c^{p \times q}$, the Kronecker product of $\boldsymbol{A}$ is a block matrix, which is defined as follows:

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix} \in c^{mp \times nq} \tag{6}$$

Moreover, each 32 × 32 pixel region in the input image corresponds to a 1024-dimensional vector in the output of ResBlock-3, i.e., $\boldsymbol{O}_{\text{ResBlock-3}}$. Accordingly, the size of $\boldsymbol{O}_{\text{ResBlock-3}}$ is $(N/32) \times (N/32) \times 1024$ pixels. Because each feature vector in $\boldsymbol{O}_{\text{ResBlock-3}}$ is shared by sixteen defined feature points, the Kronecker product is performed on the output of $\boldsymbol{O}_{\text{ResBlock-2}}$ to obtain the third feature map, $\boldsymbol{F}_3$:

$$\boldsymbol{F}_3 = \boldsymbol{O}_{\text{ResBlock-3}} \otimes \boldsymbol{I}_{4 \times 4 \times 1} \tag{7}$$

Then, three output feature maps, namely $\boldsymbol{F}_1$, $\boldsymbol{F}_2$, and $\boldsymbol{F}_3$, produced by the ResNet-50 feature extraction network are concatenated into one feature description map $\boldsymbol{F}$. It contains the information of multiple layers, and its size is $(N/8) \times (N/8) \times 2048$ pixels. Every 2048-dimensional component in $\boldsymbol{F}$ corresponds to an 8 × 8 pixel region of the input image.

As can be seen, the size of three feature maps of the ResNet-50 network applied in feature description vector construction is $(N/8) \times (N/8) \times 512$, $(N/16) \times (N/16) \times 512$ and $(N/32) \times (N/32) \times 1024$, respectively. Therefore, it is necessary to up-sample $\boldsymbol{O}_{\text{ResBlock-2}}$ and $\boldsymbol{O}_{\text{ResBlock-3}}$. Two types of up-sampling are adopted in this work for the purpose of comparison. One is to combine three feature components of a feature point into one description vector by the Kronecker product, as mentioned before. The other is up-sampling through bilinear interpolation [25]. It is the expansion of linear interpolation for two-dimensional rectangular grids. It performs interpolating on bivariate functions, and its essence is one-dimensional linear interpolation, respectively, in two directions.

2.2.2. Feature Matching

Feature description vectors should be normalized before they are exploited in feature matching. In this work, defined feature points are matched by using the Euclidean distance as the similarity measure, i.e., comparing the geometric distance between feature descriptors of feature points.

$$d[x_i, y_j] = d[\boldsymbol{F}(x_i), \boldsymbol{F}(y_j)] = \sqrt{\sum_{k=1}^{2.048} \left[f_k(x_i) - f_k(y_j)\right]^2} \tag{8}$$

where $d(x_i, y_j)$, for i and j = 1, 2, 3, . . . , is the distance between the i-th point x_i in an image to be registered and the j-th point y_j in its reference image. $\boldsymbol{F}(x_i)$ and $\boldsymbol{F}(y_j)$ represent the feature description vectors of the two points, respectively. Every component of feature description vectors is denoted by $f(\cdot)$. Given that point y_j in the reference image—the point in the image to be registered that makes the similarity measure $d(x_i, y_j)$ minimum—can be regarded as the associated feature point of y_j. After obtaining feature point mappings between the inputted images, these point coordinates should be restored to the original images for screening false matches and fitting a transform model.

2.3. False Match Elimination and Transform Model Fitting

In this work, the PROSAC algorithm [26] is utilized to sift false matches. Compared with uniform sampling from a set of matched point pairs by RANSAC [9], the PROSAC algorithm sorts all the matched point pairs according to a similarity metric. Then it samples from an increasing optimal set of matched point pairs. This method cannot only save calculation costs, but also improve operation speed. The points in a sample set are

reordered in advance. The inner points for effectively estimating model are upper in the rankings, while the outer points that have negative influences on influence are lower ranked. Therefore, fitting models are fulfilled through sampling from the upper ranked point pairs, which reduces randomness of the algorithm and enhances the success rate of obtaining a proper model. The accuracy of image registration is further improved.

In addition, the registration method proposed in this work is primarily aimed at scene mosaicking and panorama generation. Thus, only those UAV images with small differences in shooting angles are taken into consideration. In this case, it can be believed that the requirements of homography transform are satisfied approximately. Therefore, a homography matrix is employed in the final registration transform for UAV images, since it is more suitable for two-dimensional content matching and scene expansion. If the shooting angles of images to be registered are quite different, this kind of registration problem should be solved through stereo matching approaches.

3. Results

In the experimental part of this work, the deep feature extraction methods were investigated for image registration in detail. Then, in this paper, our proposed registration method is compared with existing traditional registration algorithms and a state-of-art registration algorithm based on deep learning. The experiments were performed on UAV visible-light images from different sensors and scenes. The experimental results indicate that our proposed algorithm can provide higher registration accuracy than the other algorithms. The hardware platform for experiments is configured with an Intel Core i5-4590k processor at 3.5 GHz main frequency and a 16 GB RAM. The software environment is built by the 64-bit Windows10 operating system and tensorflow-1.13.1 deep learning framework. The Python version is 3.7.0.

3.1. Experimental UAV Images

In order to verify the stability and applicability of registration algorithms, the UAV visible-light image pairs were separately collected from five different typical scenes, including urban, buildings, roads (by a river), farmlands and forests. Among them, the city and building scene images are taken from the drone image dataset downloaded from the ISPRS official website [10]. This dataset was built in 2014 and contains a total of 26 high-altitude drone images, 1000 images taken close to the ground and 1000 ground shots. In this work, two high-altitude images and two near-ground images were selected from the dataset, including the urban and building scenes. The size of these images is 2044 $\times$ 1533 and 3000 $\times$ 2000 pixels, respectively. The road scene image pair is obtained from the UAV sample images provided by the Pix4Dmapper software. This testing dataset contains 13 drone images taken near the ground in 2013. The size of these images is 2000 $\times$ 1500 pixels. The UAV image pair about farmland scene was taken by a Parrot Sequoia camera with 5 mm focal length in September 2017 at Dayi County, Sichuan Province, China. The sensor resolution is 2404 $\times$ 1728 pixels, and the shooting height is 80 m. The near-ground image pair of a forest scene is acquired in June 2019 by a Zenmuse Z30 camera, whose minimum focal length is 10 mm. The image size is 1920 $\times$ 1080 pixels and the camera height is 152 m. The observation location is in Wusufo Mountain National Forest Park, Xinjiang Uygur Autonomous Region, China. All the UAV images used in our experiments are presented in Figure 4.

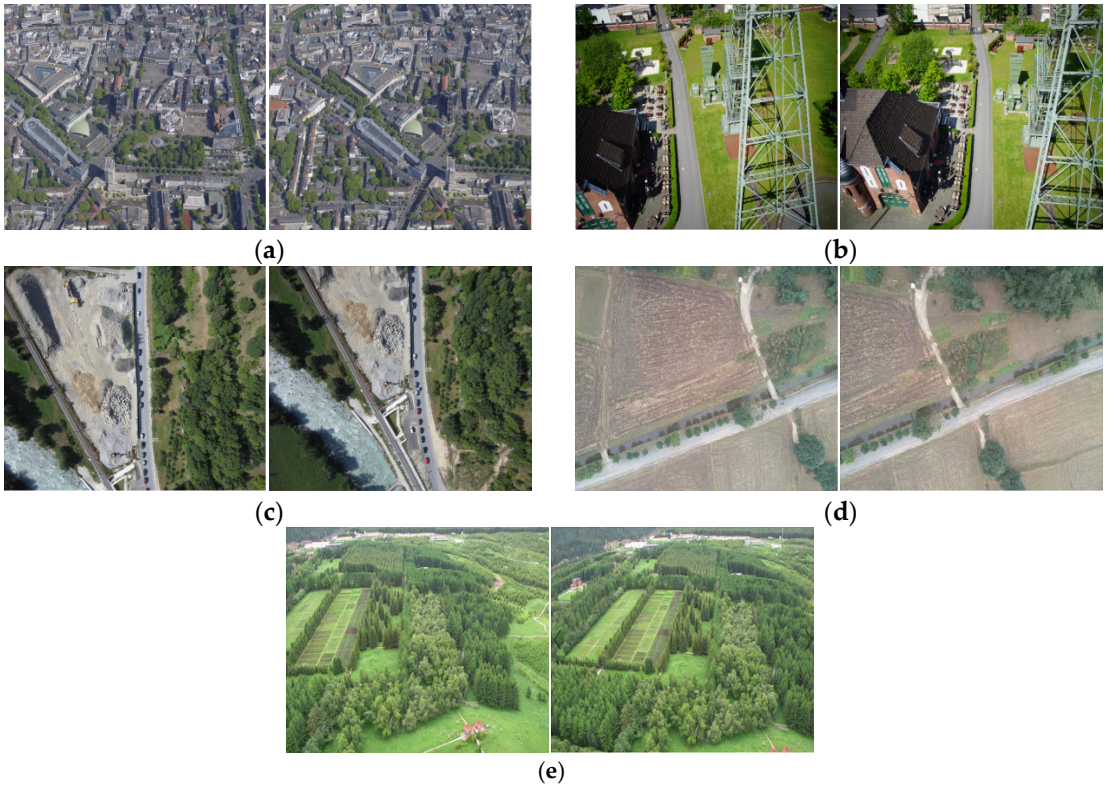
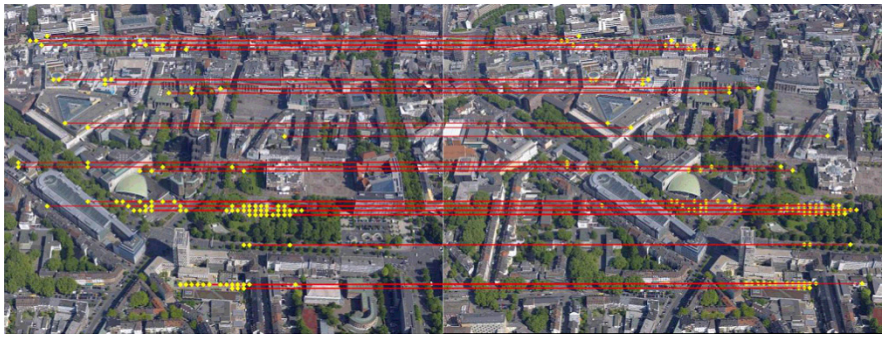


Figure 4. UAV visible-light images about different scenes: (**a**) Urban (2044 × 1533 pixels), (**b**) Buildings (3000 × 2000 pixels), (**c**) Roads (2000 × 1500 pixels), (**d**) Farmlands (2404 × 1728 pixels) and (**e**) Forests (1920 × 1080 pixels).

3.2. Visual Evaluation of Registration Results

Taking the urban scene as an example, two feature extraction methods based on deep neural networks, VGG-16 [23] and ResNet-50, are compared. In addition, distance weighting, bilinear interpolation and the Kronecker product are examined in building deep residual feature vectors of ResNet-50 network. The matched feature point pairs obtained by these methods are presented in Figure 5. DResNet-50 represents a deep residual registration method by feature distance weighting. It is similar to the method of VGG-16. BResNet-50 represents the deep residual registration method by using bilinear interpolation to realize up-sampling. KResNet-50 represents our deep residual registration method by using the Kronecker product to integrate feature vectors.

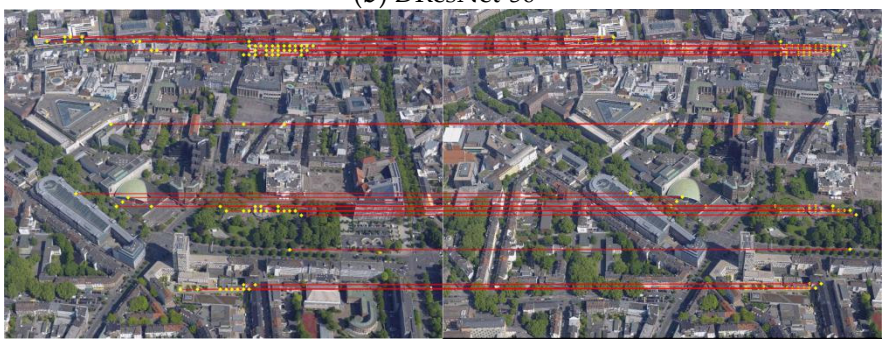
From the above five pairs of feature point matching images, it can be clearly seen that three methods based on the deep residual neural network can obtain more evenly distributed matched feature points in image overlap areas than the method based on VGG-16. These results will lead to improvements in the final registration accuracy.



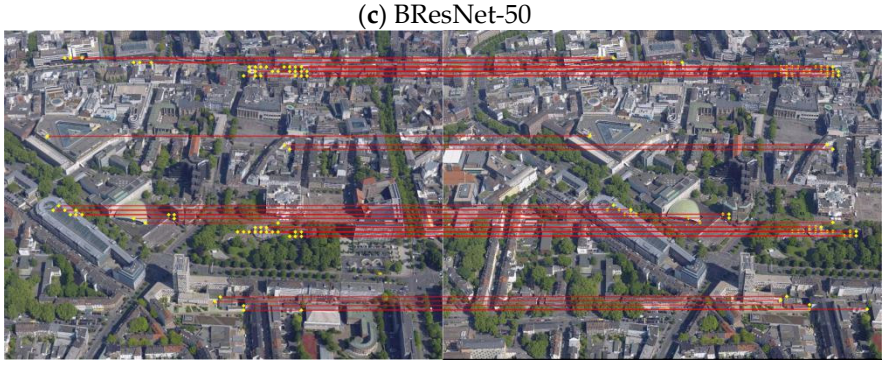
(**a**) VGG-16

Figure 5. *Cont.*

(**b**) DResNet-50



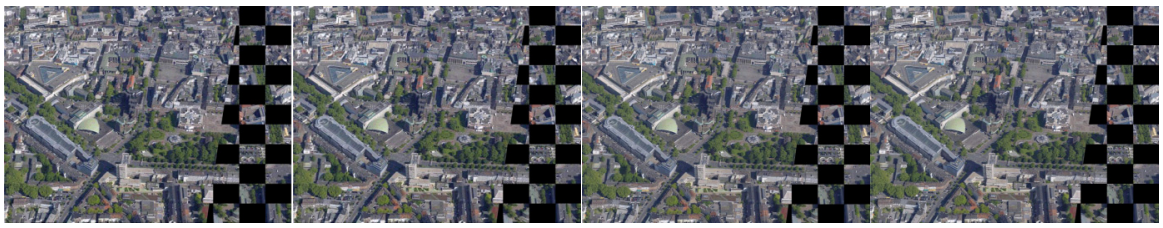
(**c**) BResNet-50



(**d**) KResNet-50

Figure 5. The matched point pairs for the urban scene image: (**a**) VGG-16, (**b**) DResNet-50, (**c**) BResNet-50 and (**d**) KResNet-50.

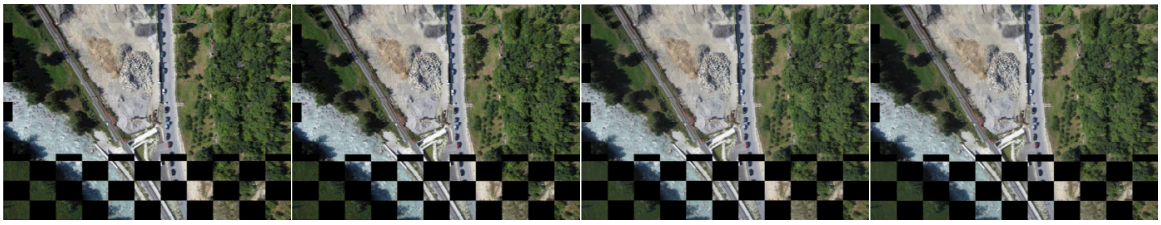
Furthermore, checkerboard mosaicked images for the urban scene obtained by different registration methods are displayed in Figure 6. Generally, a checkerboard mosaicked image is generated by alternately piecing blocks from a registered image and its reference image. In this manner, alignment details between the registered image and its reference image can be manifested. Some details of theses checkerboard mosaicked images are given in Figure 7. Because there are a little visual differences, no detailed comparison of the mosaicked images about the road scene is presented.



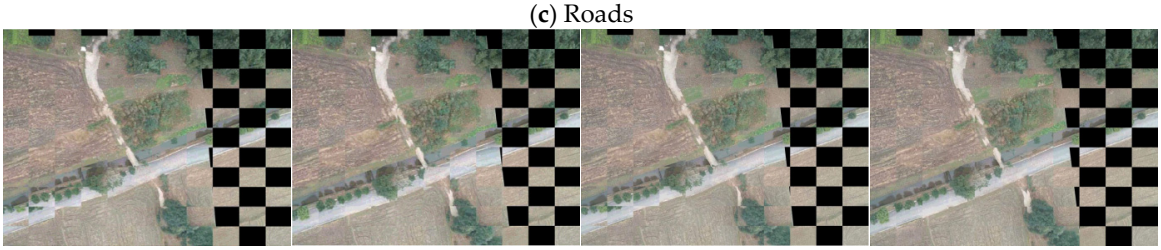
(**a**) Urban

Figure 6. *Cont.*

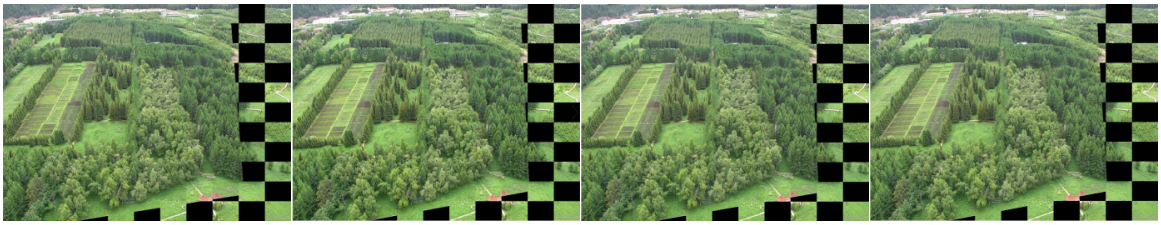
(**b**) Buildings



(**c**) Roads



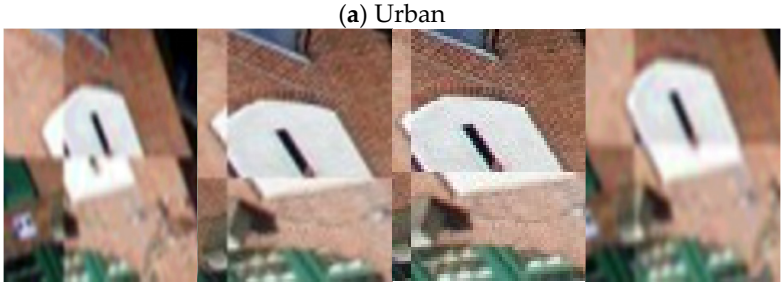
(**d**) Farmlands



(**e**) Forests

Figure 6. Checkerboard mosaicked images of five scenes obtained by four registration methods, and from left to right, they are VGG-16, DResNet-50, BResNet-50 and KResNet-50: (**a**) Urban, (**b**) Buildings, (**c**) Roads, (**d**) Farmlands and (**e**) Forests.

(**a**) Urban

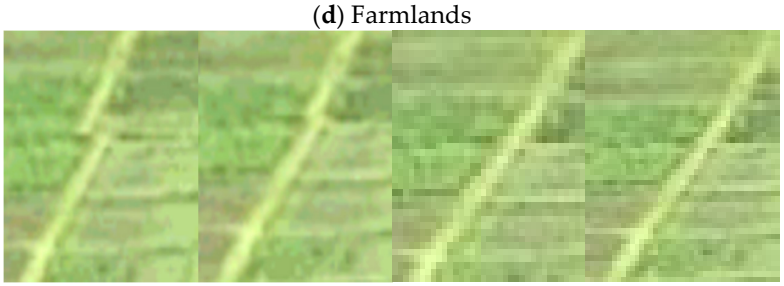


(**b**) Buildings1

Figure 7. *Cont*.

(**c**) Buildings2

(**d**) Farmlands



(**e**) Forests

Figure 7. Some details in the checkerboard mosaicked images of five scenes obtained by four registration methods, and from left to right, they are VGG-16, DResNet-50, BResNet-50 and KResNet-50: (**a**) Urban, (**b**) Buildings1, (**c**) Buildings2, (**d**) Farmlands and (**e**) Forests.

As shown in Figure 6, the matching effects of using our method based on the ResNet-50 features are better than the image registration method based on the VGG-16 features. This reason is that ResNet-50 has a deeper network structure and can generate feature vectors with higher dimensions, which is beneficial for distinguishing false and correct matched points and making mosaicking results more accurate. Moreover, in terms of details, it can be seen from Figure 7 that the Kronecker product integration method outperforms the distance weighting method and the bilinear interpolation method. The bilinear interpolation fusion method has the lowest accuracy. One reason is that, as for the distance weighting method, the similarity between the feature vectors, extracted in three different scales, are calculated respectively. The obtained results are weighted to get the final feature similarity for matching feature points. The processing method may enlarge feature distances improperly and cannot well represent the real similarity between feature points. The other reason is that the bilinear interpolation of feature vectors will assign inappropriate estimation values for the low-scale feature vectors, which leads to the increase of registration error. The Kronecker product method combines the residual feature vectors according to the relationship between them and feature points. It preserves the information of feature vectors to the most extent and improves image registration quality.

3.3. Quantitative Comparison of Registration Results

In terms of registration accuracy, besides the methods based on deep neural networks, the root mean square error (RMSE) results for the five scenes gained by other common image registration algorithms, including ORB [27], SIFT [28], SURF [29], KAZE [30], AKAZE [31], CFOG [32] and KNN + TAR [33], are listed in Table 1. The running time of the ORB algorithm is proportional to the number of feature points. The more feature points

that are required, the longer the running time of the algorithm is. Therefore, compromising running time and accuracy, the number of feature points for ORB is pre-set at 1000. Moreover, CFOG and KNN + TAR algorithms are implemented on the Matlab software platform with lower code efficiency. In the community of registration, especially involving point registration, RMSE can be expressed in the following form:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\|T(y_i, \boldsymbol{\theta}) - x_i\|)^2} \tag{9}$$

where x_i and y_i, for $i = 1, 2, 3, \ldots$, respectively represent the matching point pairs from the image to be registered and the reference registration image. Suppose that there is a total of n pairs. T is a transform model, $\boldsymbol{\theta}$ is the model parameter vector and $\|\cdot\|$ represented is the Euclidean distance between the two points. Generally, the smaller the value of RMSE, the higher the registration accuracy.

Table 1. Accuracy comparison of different registration algorithms (pixels).

Methods	Urban	Roads	Buildings	Farmlands	Forests
ORB [27]	1.32818	1.35295	1.32732	1.20049	1.28705
SIFT [28]	1.23216	1.18053	1.17576	1.37352	1.26922
SURF [29]	1.12424	1.26695	1.29178	1.39047	1.33442
KAZE [30]	1.18462	1.29448	1.21727	1.26681	1.22871
AKAZE [31]	1.02061	1.15461	1.11633	1.16056	1.23265
CFOG [32]	33.9525	37.9518	39.1872	33.9503	35.7468
KNN + TAR [33]	1.40850	2.50624	5.96340	1.88389	6.99030
VGG-16 [23]	1.07819	1.01689	1.02182	1.06978	1.02238
DResNet-50	0.98294	0.96423	1.01685	0.95157	0.93103
BResNet-50	0.99255	1.02085	1.06765	1.02273	0.91334
KResNet-50	0.94289	0.97997	0.99376	0.92051	0.90167

As given in Table 1, it can be found that the first five methods based on point features perform well on the UAV test images of different scenes. The reason for the low accuracy of the CFOG algorithm may be that it is more suitable for heterogeneous optical image registration. KNN + TAR algorithm is unstable, and it may be more suitable for satellite-borne optical image registration. The registration methods based on deep networks all provide higher registration accuracy for different image scenes than other current algorithms. KResNet-50 can even offer subpixel accuracy for five scenes. Differing from the registration methods based on point features, the methods based on depth learning are not dependent on complex contents and detail information in image scenes. As for dealing with the UAV images with simple scenes and less detail information, such as farmlands and forests, they still exhibit good registration performances. The primary reason is that, in deep-learning-based methods, the number of feature points is not directly determined by visible features of input images, but by their sizes. Many features used in registration are deep features of images. Moreover, compared with the existing method based on traditional convolutional neural networks, the proposed deep residual registration method can extract more effective information for registration. The reason is that, compared with ordinary CNN networks, the residual structures of ResNet-50 can effectively solve the problems about gradient vanishing, explosion and network degradation caused by the increase of network layers. It can make gradient information corresponding to defined feature points transfer smoothly in forward and back propagations. Thus, higher-level information can be effectively extracted for the construction of feature-point description vectors. Therefore, the methods based on ResNet-50 provide better accuracy in Table 1. Additionally, the time complexity comparisons of different registration algorithms for the five scenes are presented in Table 2.

Table 2. Time complexity comparison of different registration algorithms (seconds).

Methods	Urban	Roads	Buildings	Farmlands	Forests
ORB [27]	2.25200	2.32400	3.45700	1.41400	1.72700
SIFT [28]	125.25400	137.02500	163.34800	50.89900	93.88900
SURF [29]	64.02300	86.79900	139.72500	46.65200	42.79400
KAZE [30]	128.40800	180.03100	194.02800	72.71200	84.77400
AKAZE [31]	101.05300	144.14600	117.05400	35.76800	103.44700
CFOG [32]	4.38921	4.56470	4.41006	4.36076	4.36785
KNN + TAR [33]	24.98238	7.96253	12.62796	2.87249	28.61729
VGG-16 [23]	179.06431	94.41617	187.56367	116.60120	193.42121
DResNet-50	205.24510	102.79676	209.37979	130.89990	200.82188
BResNet-50	223.31729	118.71738	217.50212	143.52165	222.49254
KResNet-50	219.79922	114.83660	225.45332	142.06583	222.08461

From Table 2, it can be seen that the ORB algorithm has obvious advantages in running time. The reason is that, due to high resolutions of UAV images, other algorithms should find feature points as much as possible. Hence, more time was consumed in the processes of feature-point detection and matching, whereas it can be observed that the registration accuracy of all other algorithms is almost higher than that of ORB, except CFOG and KNN + TAR. This result is also because of the number of feature points. Compared with the VGG-16 registration algorithm, the time increments of the registration algorithm based on deep residual network are about 10 to 20 s, which is acceptable for registration performance improvement.

4. Discussion and Conclusions

In this work, an automatic registration method for drone images based on the deep residual network feature was proposed. The method needs no additional training and does not depend on specific contents of images. It takes the center point of each 8×8 pixel region of an input image as a feature point and constructs multi-scale feature description vectors of the feature point from the output vector of three residual network layers by the Kronecker product. For matching feature points, the PROSAC algorithm is utilized to sifting outliers and fit a geometric transform model. The experimental results for UAV images from different scenes indicate that combining deep residual features and PROSAC can fulfill high accurate, even subpixel, registration. Compared with existing state-of-the-art registration algorithms, it is manifested that the proposed image registration method based on deep residual features exhibits remarkable performance enhancements.

Although deep residual network features can describe images in-depth, these features are of high dimensions that are proportional to the size of input images. Therefore, our method has no advantage in running time, and subsequent research studies can be focused on reducing its computation complexity. In addition, the deep feature extraction network used in this work is a pre-trained model trained by ImageNet. It is more suitable for natural images. In the future, some more appropriate UAV image datasets can be adopted to tune the network model weights finely, so that the deep residual network can extract more distinctive features from UAV images to ameliorate registration results. Meanwhile, more false match sifting methods can be adopted in the future, or deep learning can be directly utilized to fit transformation parameters, so as to realize an end-to-end registration framework. The effects of acquisition environmental changes, such as illumination and shadows, on drone images also need to be further studied in detail. Moreover, the registration of multi-source heterogeneous images, such as infrared, near-infrared and multispectral, is also a problem that needs to be solved.

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

Funding: This research was funded by the Science and Technology Program of Sichuan, grant number 2020YFG0240 and 2020YFG0055, the Science and Technology Program of Hebei, grant number 20355901D and 19255901D.

Acknowledgments: We would like to thank the anonymous reviewers for their valuable and helpful comments, which substantially improved this paper. At last, we also would also like to thank all of the editors for their professional advice and help.

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

References

1. Li, W.; Li, C.; Wang, F. Research on UAV image registration based on SIFT algorithm acceleration. In Proceedings of the 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019.
2. Yang, C. A Compilation of UAV applications for precision agriculture. *Smart Agric.* **2020**, *2*, 1–22.
3. Tsouros, D.C.; Bibi, S.; Sarigiannidis, P.G. A review on UAV-based applications for precision agriculture. *Information* **2019**, *10*, 349. [CrossRef]
4. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision (ICCV), Kerkyra, Greece, 20–27 September 1999.
5. Zhang, M.L.; Li, S.; Yu, F.; Tian, X. Image fusion employing adaptive spectral-spatial gradient sparse regularization in UAV remote sensing. *Signal. Process.* **2020**, *170*, 107434. [CrossRef]
6. Jeong, D.M.; Kim, J.H.; Lee, Y.W.; Kim, B.G. Robust weighted keypoint matching algorithm for image retrieval. In Proceedings of the 2nd International Conference on Video and Image Processing (ICVIP 2018), Hong Kong, China, 29–31 December 2018.
7. Wang, C.Y.; Chen, J.B.; Chen, J.S.; Yue, A.Z.; He, D.X.; Huang, Q.Q.; Yi Zhang, Y. Unmanned aerial vehicle oblique image registration using an ASIFT-based matching method. *J. Appl. Remote. Sens.* **2018**, *12*, 025002. [CrossRef]
8. Bay, H.; Tuytelaars, T.; Gool, L.V. SURF: Speeded up robust features. *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
9. Hossein-Nejad, Z.; Nasri, M. Image registration based on SIFT features and adaptive RANSAC transform. In Proceedings of the 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, India, 6–8 April 2016.
10. Nex, F.; Gerke, M.; Remondino, F.; Przybilla, H.J.; Bäumker, M.; Zurhorst, A. ISPRS benchmark for MultiPlatform photogrammetry. *ISPRS Ann.* **2015**, *II-3/W4*, 135–142.
11. Yu, R.; Yang, Y.; Yang, K. Small UAV based multi-viewpoint image registration for extracting the information of cultivated land in the hills and mountains. In Proceedings of the 26th International Conference on Geoinformatics, Kunming, China, 28–30 June 2018.
12. Fernandez, P.; Bartoli, A.; Davison, A. KAZE features. In Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012.
13. Gu, J.; Wang, Z.; Kuen, J. Recent Advances in Convolutional Neural Networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
14. Meng, L.; Zhou, J.; Liu, S.; Ding, L. Investigation and evaluation of algorithms for unmanned aerial vehicle multispectral image registration. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102403. [CrossRef]
15. Fischer, P.; Dosovitskiy, A.; Brox, T. Descriptor matching with convolutional neural networks: A comparison to SIFT. *Comput. Sci.* **2014**, *4*, 678–694.
16. Nassar, A.; Amer, K.; ElHakim, R.; ElHelw, M. A deep CNN-based framework for enhanced aerial imagery registration with applications to UAV geolocalization. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018.
17. Nguyen, T.; Chen, S.W.; Shivakumar, S.S.; Taylor, C.J.; Kumar, V. Unsupervised deep homography: A fast and robust homography estimation model. *IEEE Robot. Autom. Lett.* **2018**, *3*, 2346–2353. [CrossRef]
18. Ye, F.; Su, Y.; Xiao, H. Remote sensing image registration using convolutional neural network features. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 232–236. [CrossRef]
19. Wang, X.; Zeng, W.; Yang, X. Bi-channel image registration and deep-learning segmentation (BIRDS) for efficient, versatile 3D mapping of mouse brain. *Nat. Libra Medic.* **2021**, *10*, e63455.
20. Zhang, R.; Xu, F.; Yu, H.; Yang, W.; Li, H.C. Edge-driven object matching for UAV images and satellite SAR images. In Proceedings of the 2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020.
21. Nazib, R.A.; Moh, S. Energy-efficient and fast data collection in UAV-aided wireless sensor networks for Hilly Ter-Rains. *IEEE Access* **2021**, *9*, 23168–23190. [CrossRef]
22. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.

23. Bae, W.; Yoo, J.; Ye, J.C. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017.
24. Yang, Z.; Dan, T.; Yang, Y. Multi-temporal remote sensing image registration using deep convolutional features. *IEEE Access* **2018**, *6*, 38544–38555. [CrossRef]
25. Ye, H.; Su, K.; Huang, S. Image enhancement method based on bilinear interpolating and wavelet transform. In Proceedings of the 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 March 2021.
26. Chum, O.; Matas, J. Matching with PROSAC—Progressive sample consensus. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 21–23 September 2005.
27. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011.
28. López, A.; Jurado, J.M.; Ogayar, C.J.; Feito, F.R. A framework for registering UAV-based imagery for crop-tracking in precision agriculture. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *97*, 102274. [CrossRef]
29. Jhan, J.; Rau, J. A generalized tool for accurate and efficient image registration of UAV multi-lens multispectral cameras by N-SURF matching. *IEEE J. Sel. Top Appl. Earth Obs. Remote Sens.* **2021**, *14*, 6353–6362. [CrossRef]
30. Mohamed, K.; Adel, H.; Raphael, C. Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach. *Comput. Electron. Agric.* **2020**, *174*, 105446.
31. Yan, J.; Wang, Z.; Wang, S. Real-time tracking of deformable objects based on combined matching-and-tracking. *J. Electron. Imaging* **2016**, *25*, 023019. [CrossRef]
32. Ye, Y.; Bruzzone, L.; Shan, J.; Bovolo, F.; Zhu, Q. Fast and robust matching for multimodal remote sensing image registration. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9059–9070. [CrossRef]
33. Hou, X.; Gao, Q.; Wang, R.; Luo, X. Satellite-borne optical remote sensing image registration based on point features. *Sensors* **2021**, *21*, 2695. [CrossRef] [PubMed]