remote sensing

MDPI

Article

LLFE: A Novel Learning Local Features Extraction for UAV Navigation Based on Infrared Aerial Image and Satellite Reference Image Matching

Xupei Zhang, Zhanzhuang He, Zhong Ma *, Zhongxi Wang and Li Wang

Xi'an Microelectronics Technology Institute, Xi'an 710065, China; zxp771tuantuan@163.com (X.Z.); hzz771@163.com (Z.H.); zxw0919@163.com (Z.W.); wangli009g@163.com (L.W.)
* Correspondence: mazhong@mail.com; Tel.: +86-1357-181-4259

Abstract: Local features extraction is a crucial technology for image matching navigation of an unmanned aerial vehicle (UAV), where it aims to accurately and robustly match a real-time image and a geo-referenced image to obtain the position update information of the UAV. However, it is a challenging task due to the inconsistent image capture conditions, which will lead to extreme appearance changes, especially the different imaging principle between an infrared image and RGB image. In addition, the sparsity and labeling complexity of existing public datasets hinder the development of learning-based methods in this research area. This paper proposes a novel learning local features extraction method, which uses local features extracted by deep neural network to find the correspondence features on the satellite RGB reference image and real-time infrared image. First, we propose a single convolution neural network that simultaneously extracts dense local features and their corresponding descriptors. This network combines the advantages of a high repeatability local feature detector and high reliability local feature descriptors to match the reference image and real-time image with extreme appearance changes. Second, to make full use of the sparse dataset, an iterative training scheme is proposed to automatically generate the high-quality corresponding features for algorithm training. During the scheme, the dense correspondences are automatically extracted, and the geometric constraints are added to continuously improve the quality of them. With these improvements, the proposed method achieves state-of-the-art performance for infrared aerial (UAV captured) image and satellite reference image, which shows 4–6% performance improvements in precision, recall, and F1-score, compared to the other methods. Moreover, the applied experiment results show its potential and effectiveness on localization for UAVs navigation and trajectory reconstruction application.

Keywords: image feature extraction; scene matching; visible light and infrared image; UAV and satellite imagery; UAV vision-based navigation

check for updates

Citation: Zhang, X.; He, Z.; Ma, Z.; Wang, Z.; Wang, L. LLFE: A Novel Learning Local Features Extraction for UAV Navigation Based on Infrared Aerial Image and Satellite Reference Image Matching. *Remote Sens.* **2021**, *13*, 4618. https://doi.org/10.3390/rs13224618

Academic Editors: Giancarmine Fasano and Roberto Opromolla

Received: 29 September 2021
Accepted: 13 November 2021
Published: 16 November 2021

1. Introduction

As a normal method of navigation and positioning, the GPS/INS integrated navigation system has been widely used for precise localization of UAVs. However, this system is not always available or applicable, owing to signal interference, cost, or power-consuming limitations in real application scenarios [1]. Therefore, a new low-cost navigation and positioning technique which can be robustly applied in a GPS denied environment must be considered. The process of seeking the same scene in different images through the consistency of image features, structure, and content is usually known as image matching. For decades, it has been one of the crucial techniques in various applied fields, including vision-based navigation of UAVs [2], geometric alignment [3], precise localization [4], and automatic landing and takeoff [5].

Figure 1 shows a sketch of the UAV navigation technique based on image matching. In general, the pre-acquired satellite images with real geographic labeling information are

stored in the onboard computer of the UAV. Thus, every point (pixel) such as $P_R(X, Y)$ on the center of the satellite image has the corresponding location information (such as longitude and latitude). When the UAV is constantly capturing scenes on the ground, the projection points to the spatial points in the real-world which, on the aerial image and satellite image, will satisfy the camera projection transformation model [6]. In other words, we can always find corresponding points $P_A(X, Y)$ on the real-time aerial image to represent the point $P_R(X, Y)$. Thus, the UAV location information can be obtained through the camera projection transformation model and the matched local feature points are the key to calculating the projection transformation model [7–9]. Although the image matching methods for vision-based navigation of a UAV, which is based on visible light images as the referenced and real time aerial data, has been successfully applied in cruise missile and terminal guidance missile [10], most of the commercial optical cameras are vulnerable to adverse weather or illumination conditions and cannot obtain ideal high-quality images in many cases. An infrared imaging sensor is not affected by natural factors such as fog, night, and the other limited visibility environments, and it is capable of imaging all day and all weather. Therefore, it has become one of the research hotspots to develop an image matching algorithm based on infrared images as the real-time data and visible light images as the referenced data for precise navigation and guidance technique for UAVs. On this basis, the image matching based on an infrared aerial image and satellite reference image aims to acquire the region image near the flight or target area through the infrared camera on the UAV. It will then match the stored reference satellite RGB image to acquire the projection transformation between the satellite image and infrared aerial image, and through the projection model the position data of the UAV will be obtained. However, the real-time image and reference image are captured in different seasons, at different times, or from different viewpoints, which will cause severe geometric distortion or illumination change or occlusion. Apart from the issues mentioned above, image matching on infrared image and RGB image faces additional challenges:

(1) Even in the same scenes, the different image-forming principles between different types of cameras may cause the same content to be represented by different intensity values, which means that the images from an infrared camera and RGB camera have extreme appearance changes (shown in Figure 2a,b). The poor consistency makes it difficult to find the correspondences based on the traditional image features (such as intensity or gradient values [11–14]).
(2) In recent years, deep learning techniques have shown the great ability of feature representation in scene matching and other computer vision tasks [15–17], which benefits from the explosive growth in image dataset utilization. However, the image datasets for scene matching are almost always from the common cameras (most of them are RGB images from the visual camera). Infrared aerial image and satellite reference image datasets for image matching tasks remain scarce. This is the most significant limitation to application and performance improvement of the deep learning algorithms in this research area.
(3) Even if the infrared aerial images and satellite reference images are sufficient, the learning local feature-based methods still require a large number of labels for algorithm training. Moreover, the labels of local features are difficult to be obtained by human annotation, considering the tremendous numbers and strict requirements of annotation precision. Therefore, there is an urgent need for a method that utilizes the available unlabeled dataset through a self-labeling scheme.

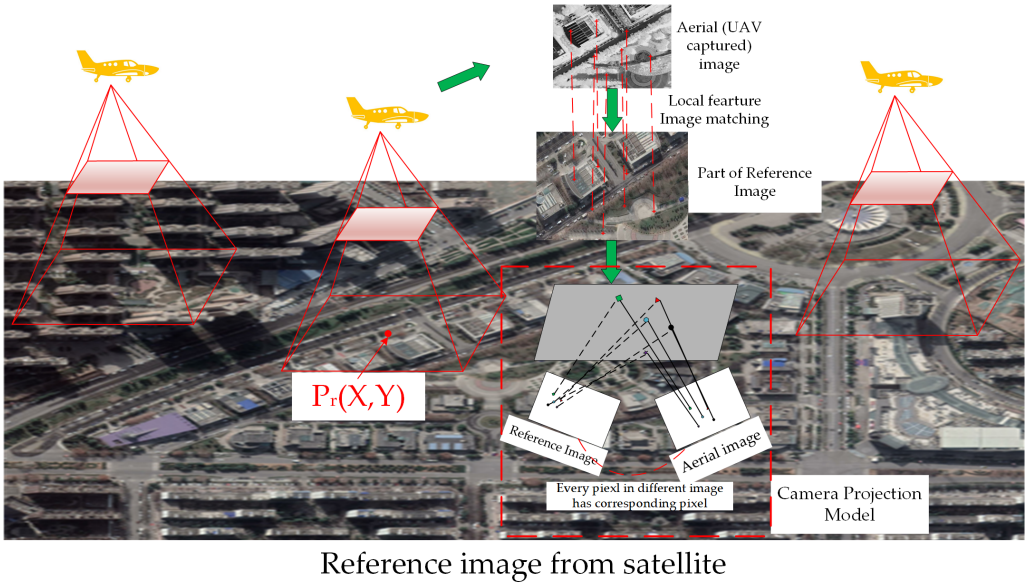

Figure 1. A sketch of the UAV navigation technique based on image matching. When the images are captured from the high altitude, it can be assumed that the spatial points in the real-world will all fall in the same plane, approximately.

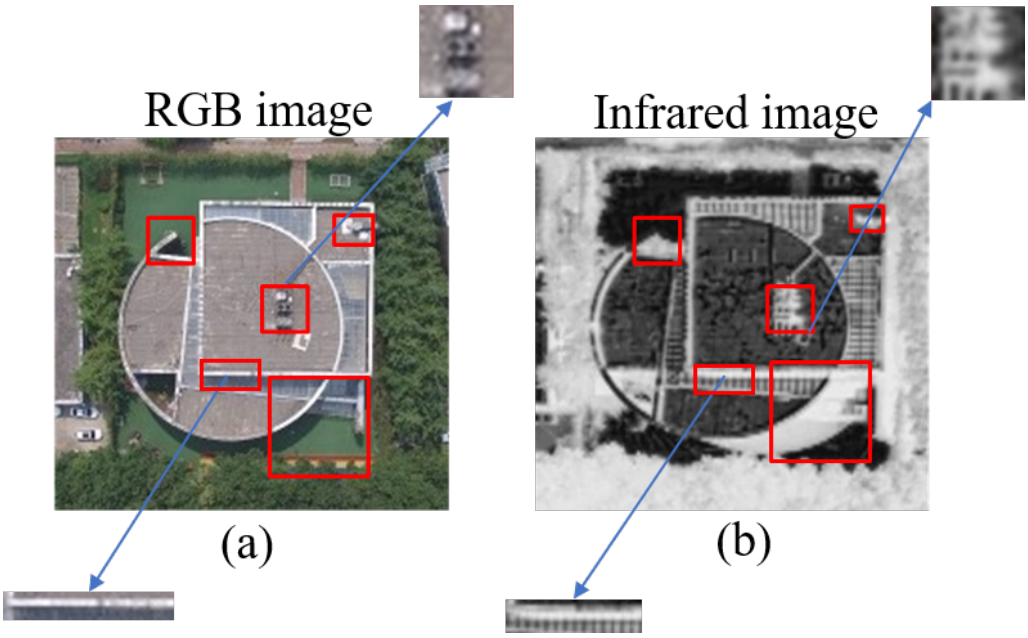

Figure 2. Different image-forming principles in sensors will cause extreme appearance changes on images. (**a**) is the RGB image sample and (**b**) is the infrared image sample.

Faced with these issues, this paper aims to learn the highly repeatable and distinctive local features to build robust and accurate correspondences with the limited and unlabeled visual and infrared image pairs. Therefore, we propose a learning detector and descriptor convolutional neural network (CNN) architecture with an iterative training scheme, named learning local features extraction (LLFE). The backbone of the network is inspired by [18] and used to obtain a set of feature maps from the input image pair. The feature maps are then utilized to detect subpixel-level local features and compute a dense and robust local descriptor for each local feature simultaneously. Meanwhile, we also design an iterative training scheme that can automatically generate and optimize the pseudo label for algorithm training to improve the performance of the proposed algorithm. In summary, the following contributions were made:

(1) Our proposed network was established and optimized to achieve a highly precise feature representation and location recognition. Specifically, a backbone network is designed to keep the input resolution consistent, and a novel detector branch with Softargmax is designed to obtain the local maxima in the feature map, which are interest points afterwards. With these improvements, the algorithm can achieve subpixel accuracy for local features detection, which can also improve the self-labeling precision for our iterative training scheme.
(2) A novel loss function was proposed for a robust local descriptor to process extreme appearance changes among infrared aerial images and satellite reference images. Unlike the popular loss function that only performs local optimization based on paired or ternary image patches, the novel descriptor loss considers the patches around the interest point and introduces average precision as the global metric for optimization to face the various challenging conditions in place recognition tasks, especially for infrared images and reference images captured from different platforms.
(3) Owing to the complex collection and a limited number of labeled infrared aerial images and satellite reference images, this paper introduces an iterative training scheme. In the training scheme, we first added geometric constrains using multiple view geometry principles that can autogenerate reliable pseudo-ground truth correspondence from the captured RGB-IR image pairs. Second, the sparse dataset can be reused to iteratively optimize the computing of the correspondence. Benefiting from these processes, the proposed method achieves state-of-the-art performance with the high-quality correspondence as the training pseudo-ground truth.

Combined with these improvements, the proposed method shows 2–3% performance improvements in precision, recall, and F1-score for scene matching on visual light and infrared image pairs. Moreover, the iterative training scheme can provide another 2–3% performance improvement by making full use of the limited training data. The experiment of trajectory reconstruction also shows the localization error of the proposed method is further decreased by a large margin (near 50%) compared to the other methods.

The rest of this work is structured as follows: In the next section, some related local feature extraction methods for scene matching task are presented. In Section 3, our proposed method LLFE is introduced. Section 4 provides the experimental results and analysis of our method and the other methods on the infrared aerial images and satellite reference images. Meanwhile, the UAV localization experiment for our method and the other methods is also provided in Section 4. The discussion of all the methods is provided in Section 5. Finally, conclusions are drawn in Section 6.

2. Related Work

In recent years, image matching between an infrared image and RGB image has received significant attention in the field of computer version. In the early stages, it was traditionally cast as region matching [19–21] or a handcrafted local feature matching task. However, the gray level information is mainly used in the region-based matching method, so it is not applicable to image matching between infrared and visible light where the image-forming principles have significantly changed the gray level of the images. Recent studies have shown that convolutional neural networks (CNNs) can extract global features [16,17] or high-level semantic features [22,23] for computer vision tasks, and some of them have shown a great perspective on image retrieval, saliency detection, and image segmentation. However, these global or semantic features are vulnerable for image matching when the captured images suffer from appearance variance caused by the different imaging principle, different viewpoint. Moreover, these features cannot represent the geometric information, which is the essential factor for the UAV vision-based navigation applications. However, the powerful image representations of CNNs motivate research on using the learning-based local features to replace the handcrafted features. Thus, the mainstream methods for image matching can be roughly categorized as handcrafted feature-based methods and learning

feature-based methods. We briefly review these related works and discuss the inspiration from these methods.

2.1. Handcrafted Feature-Based Paradigm

Classical image matching methods rely on numerous handcrafted features to establish pixel-level or subpixel-level correspondences across images. These methods started by using handcrafted features (such as blobs or corners [11,12,14]) to extract a set of interest points from the reference image. Then, local image information around each interest point (such as the difference of Gaussians (DoG) [11,12]) is used to compute the associated descriptors [11,12,14]. The obtained features and descriptors are stored in an indexing structure, such as a search tree [24]. In the scene matching step, interest points and descriptors obtained from the real-time image are used to be recognized and compared with the interest points and descriptors that are stored from the reference image. If the matching interest points are sufficient to estimate the homography or fundamental matrix using RANSAC [25], the scene is matched. The described approach has been successfully used for visible light image scene matching for many years. The most widely used algorithm is scale-invariant feature transform (SIFT) [14]. Existing research demonstrates that the listed methods work well in practice. Nevertheless, the low-level image feature information (intensity or gradient variation on images) brings limitations when the imaging conditions change drastically (e.g., day and night illumination change, weakly textured scenes, or images from infrared camera and RGB camera on different platform). In other words, the handcrafted features detection results may significantly change on infrared aerial images and satellite reference images. For example, interest points that can be visually detected in a visible image may be undetectable in an infrared image because of the distinctions in image-forming principles. Therefore, handcrafted features cannot provide stable and robust matching feature points for homography or fundamental estimation, which is crucial for scene matching. In contrast to traditional methods, deep learning methods are driven by vast amounts of data and can find more stable and robust features between infrared aerial images and satellite reference images on different source images.

2.2. Learning Feature-Based Paradigm

In contrast to traditional methods, deep learning methods are driven by vast amounts of data and can find more stable and robust features between different imaging conditions. Therefore, the majority of the learning feature-based scene matching methods started to replace the handcrafted feature-based methods [26–31]. Unlike the learned detectors [32,33] and learned descriptors [34–36] that only focus on one particular aspect (repeatable or reliable), the jointly learned descriptor and detector methods combine the repeatability of pixel-wise structures and the reliability of larger patch structures. These advantages bring significant performance in feature matching and make it that the majority of the learning feature-based methods have been developed in recent years, especially on jointly learned descriptors and detectors. LIFT [26] was the first to introduce a jointly learned descriptor and detector. The detector network finds the interest points which are then fed to the network for orientation estimation, whereas the last network creates the description of interest points. The LIFT algorithm networks do not share computed results, which renders it too slow for real-time scene matching. Super-Point [28] proposed a new network that can provide a pixel-wise local feature detection and description by the detector and descriptor branches sharing most computations, consequently accelerating the processing. It is worth mentioning that Super-Point [28] introduced a self-supervised pipeline that can artificially generate images with pseudo ground points for the algorithm training, but their pipeline cannot build the correspondence for images from infrared camera and RGB camera on different platform, which means it cannot obtain the ground truth for training the algorithm with RGB-IR images. More recently, D2-Net [30] introduced a single CNN architecture that shares all weights in the joint training process of interest point detection and description. However, their performance of precise recognition performance is worse than the other

methods, which is caused by the low-resolution feature maps. Similar to D2-Net [30], R2D2 [31] strives for a reliable descriptor, which significantly improves the robustness of the local feature matching on image queries with extreme changes. However, these methods require a clearly defined and consistently labeled dataset to achieve a good performance for scene matching. Therefore, achieving reliable and precise self-labeling ground truth and effectively using limited datasets has become a critical factor in improving the algorithm performance.Table 1 summarizes the main properties of the different types of methods.

Table 1. The main properties of handcrafted-based methods and learning-based methods.

Method	Handcrafted Feature	Learning Feature
Advantage	subpixel accuracy high speed data efficient	high-level feature repeatable and reliable discriminability and robustness
Disadvantage	low-order feature detectors require good modeling bad robust for apparent variance	require large amount of labelled data efficiency depend on structure generality only in trained regime

In summary, the handcrafted feature methods use the manual experience and knowledge a priori to find the local feature, with these local features usually based on the intensity, gradient, etc. The low-level image information is weak for appearance variances which are caused by the imaging principle or illumination condition change. Unlike the handcrafted feature methods, learning-based algorithms are data-driven methods that automatically obtain the high-level local feature extraction process and representation by automatically and directly constructing the wanted local feature structure information. Using such processes, learning feature-based algorithms with labeled datasets outperform handcrafted feature-based methods in the target recognition, segmentation, and classification tasks. However, despite this apparent success, the further application of learning feature-based methods on image matching for infrared aerial images and satellite reference images is hampered by the lack of data. Currently, the available datasets containing infrared aerial image and satellite reference image are limited and challenging to collect. Moreover, the ground truth label is difficult to obtain because human annotation on the visual and infrared image pairs is imprecise and time consuming. Therefore, the lack of labeled visual and infrared datasets impedes the training of these methods and makes the learning feature-based scene matching difficult to achieve.

3. Materials and Methods

Figure 3 shows an overview of the proposed method. The proposed method is divided into three parts: First, we applied a novel joint learning detector and descriptor convolutional neural network to extract the local features from the visible light images, called the train stage (1). It aims to let the network have the initial ability to find the local features on images and uses the trained model (premature model) to prepare the pseudo-ground truth for the next stage of training. In train stage (2), we introduce the premature model into the COLMAP [37,38] framework to obtain the dense correspondence as the pseudo-ground truth for network training on visible light and infrared images. It is worth mentioning that this stage can improve the precision of the pseudo-ground truth iteratively and the performance of our algorithm, which benefits from our iterative learning scheme. In the test stage, we use the trained proposed network to jointly predict the detector and descriptor for RGB and infrared image pairs. To combine these three stages, a novel learning-based features extraction method for RGB-IR image matching is proposed. The details are presented in the following subsections.

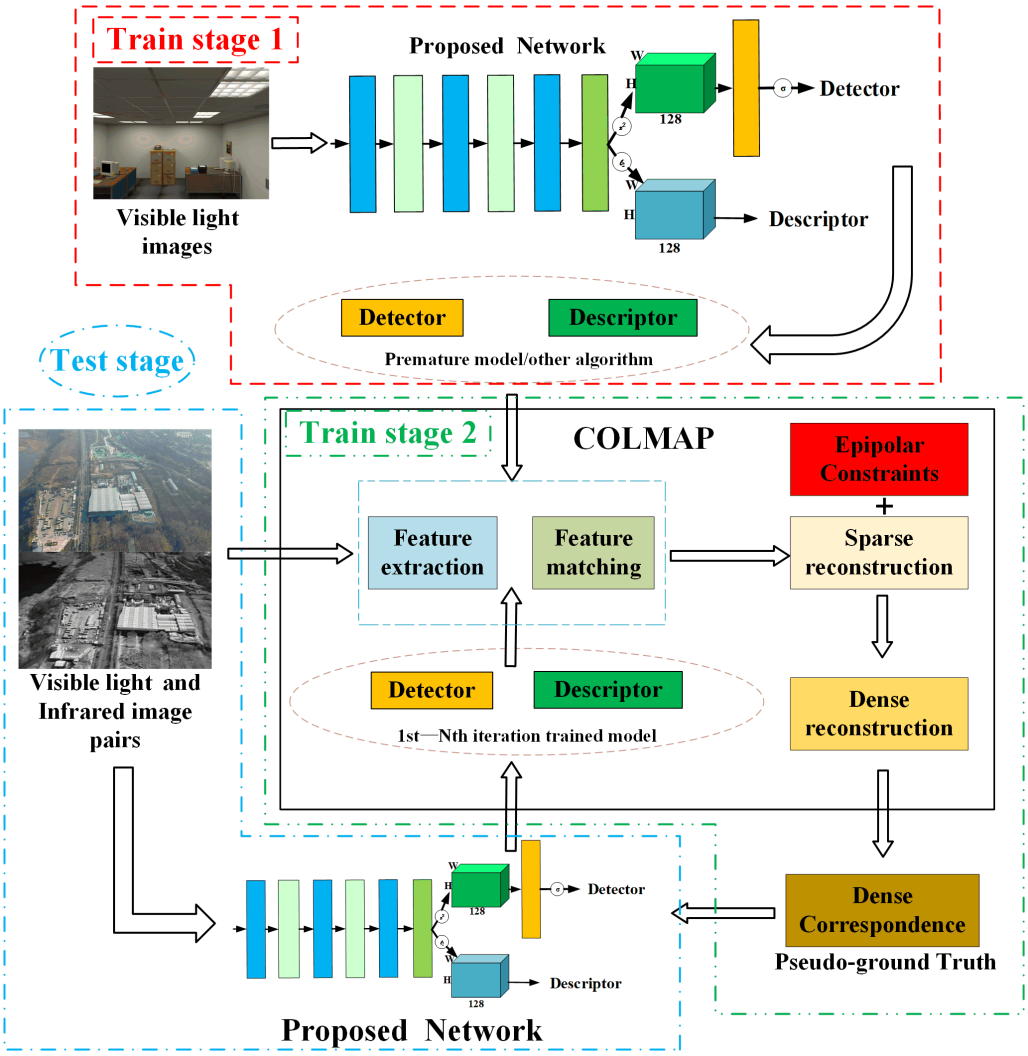

Figure 3. Overview of the proposed method.

3.1. Learning-Based Detector and Descriptor

As the crucial part of the proposed method, a joint learning detector and descriptor network architecture is shown in Figure 4. The architecture's backbone is the L2-Net [18], but there are two significant differences to improve its performance of precision recognition performance. The first difference concerns padding with zeros added after the convolutional layers (except the final one) to preserve the spatial size. The second difference relates to the use of dilated convolutions instead of subsampling to preserve the input resolution. The modified L2-Net network we used aims to predict 2 outputs for each image in the image pair (which includes I_{RGB} and I_{IR}) of size $H \times W$ (all pairs of images with at least 50% overlap). The output tensor of the backbone serves as input to two submodules. First is a L2-normalization layer that obtains a descriptor (D) $\epsilon R^{H \times W \times F}$(F = 128) for describing the structure information of the patch around each local feature (interest point) to provide the description for the local feature, and the second is an element-wise square operation followed by a 1×1 convolutional layer and a Softargmax function to extract a heatmap ($S \epsilon H \times W \times 1$) which aims to provide the local feature positions on the image.

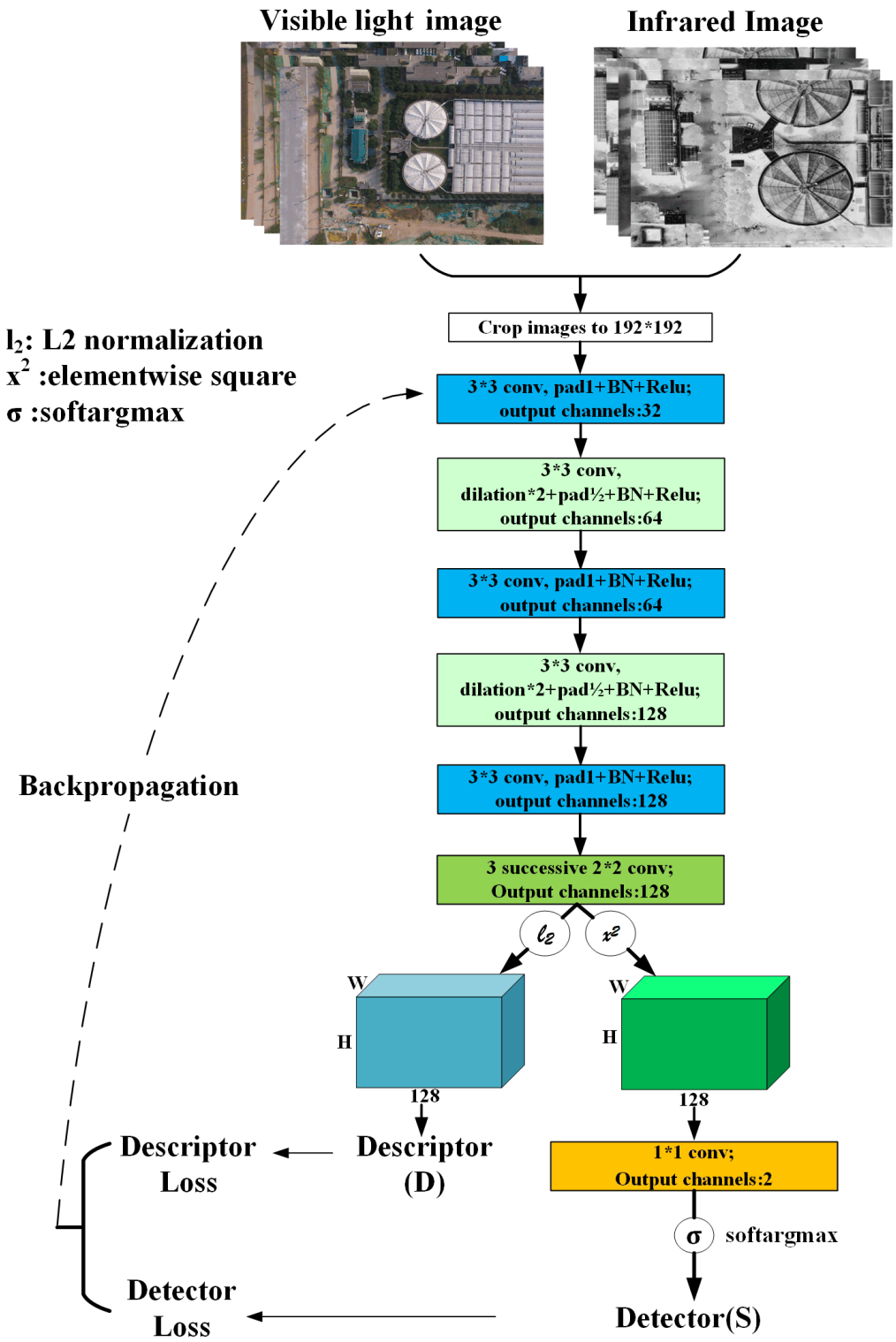

Figure 4. Overview of the network architecture.

Feature Detector: The Super-Point [28] and D2-Net [30] algorithms apply the non-maximum suppression (NMS) to obtain sparse interest points. However, NMS only provides pixel-level accuracy for interest point detection. In addition, NMS is non-differentiable. Inspired by LIFT [26], this work utilizes Softargmax to obtain the local

maxima in S as each interest point. Thus, the coordinates of each interest point and the points near the interest point on the feature map can be represented as:

$$P_{map}(x, y) = (x_{int}, y_{int}) + (\Delta x, \Delta y) \tag{1}$$

$$\Delta x = \frac{\sum_n \sum_m e^{f(x_m, y_n)} m}{\sum_n \sum_m e^{f(x_m, y_n)}}, \Delta y = \frac{\sum_n \sum_m e^{f(x_m, y_n)} n}{\sum_n \sum_m e^{f(x_m, y_n)}} \tag{2}$$

where the center pixel of the patch (the interest point) is denoted as (x_{int}, y_{int}) and denotes the value of pixel on the feature map at the coordinate (x, y). Offset on the x and y axes relative to the center point are denoted as m and n, respectively. Thus, the position of interest point on the feature map $P_{map}(x, y)$ can be updated with subpixel accuracy.

Feature Descriptor: Inspired by previous works [30,31], the descriptor is a 3D tensor $D \epsilon R^{H \times W \times F}$ representing a set of F-dimensional descriptors for each interest point. In contrast to the interest point detector that focuses on small image regions, the descriptors consider high-level structures on the larger patches around the interest points. As Figure 5 shows, when the input image is sent into the network, it will obtain the feature maps with different response value. The maximum response values in local small regions can be treated as the local features. The region (image patch) around the local feature has structure information to help the matching algorithm to locate the local feature position in a different image. For the handcrafted method, the structure information usually uses the gradient statistic, local intensity, or local intensity order statistic to describe the structure information [39]; the structure information is stable even if the image has geometric transformation. However, the low level image cues (gradient statistic, local intensity, etc.) are vulnerable when the images suffer from extreme appearance changes caused by the different imaging principles of different visual sensors. For our method (or the other deep learning methods), we use the response value statistic to describe the structure information of the image patch around the local feature. Benefiting from the great ability of high-level image cues captured by CNNs, the response values obtained by the CNN network are much more robust than the local feature description based on gradient statistic or local intensity [39]. It is also the reason that our algorithm can find the robust local feature for image matching. Thus, the descriptors can successfully match points even under the condition of substantial appearance changes, such as those in infrared aerial images and satellite reference images. Inspired by [28,30], this work applies an L2 normalization on descriptors prior to obtaining dense descriptors.

As Figure 6 shows, when the input image pair is sent into the network, a large number of regions with large response value on the S can be obtained. The learning local features are obtained by using the NMS operation on these regions with larger response value; thus, one local region can obtain one point (local feature) with the maximum response value. In other words, the network can set the threshold of response value to increase or decrease the number of the local regions to decide the number of detected local features. The rest of the work is to design the detector and descriptor loss to help the network find the robust local feature with accurate position on the image pair.

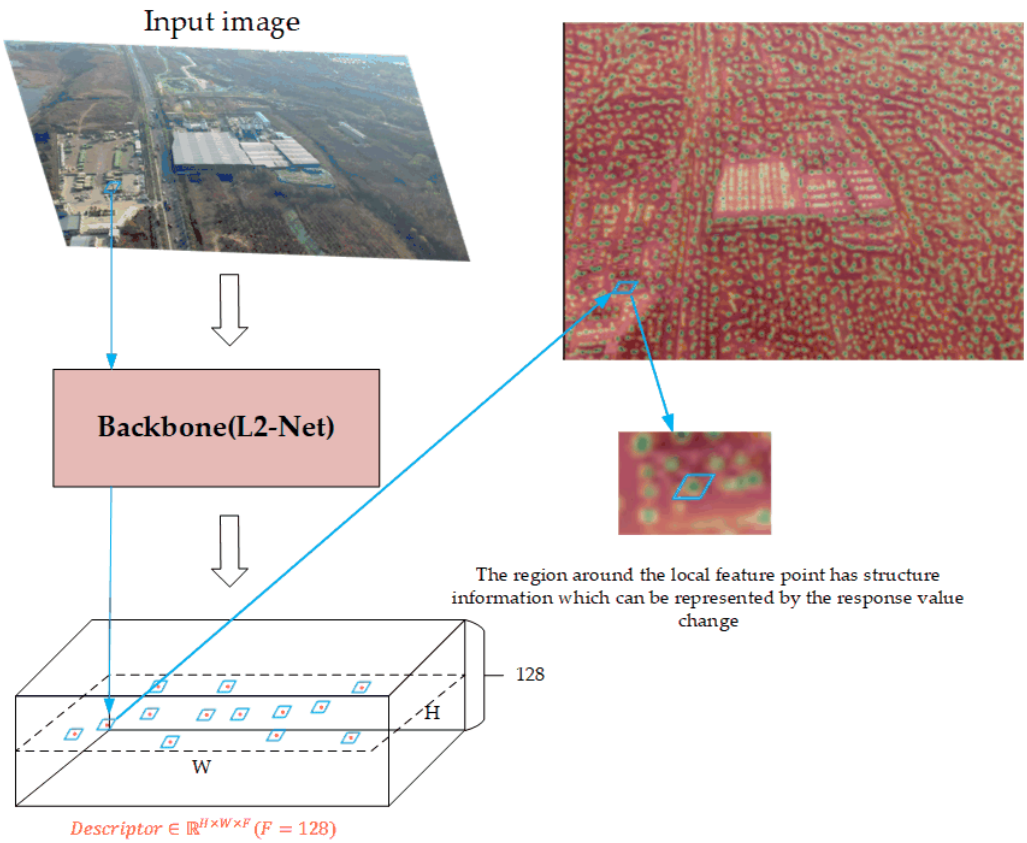

Figure 5. The description processing of the local feature on one of the input images.

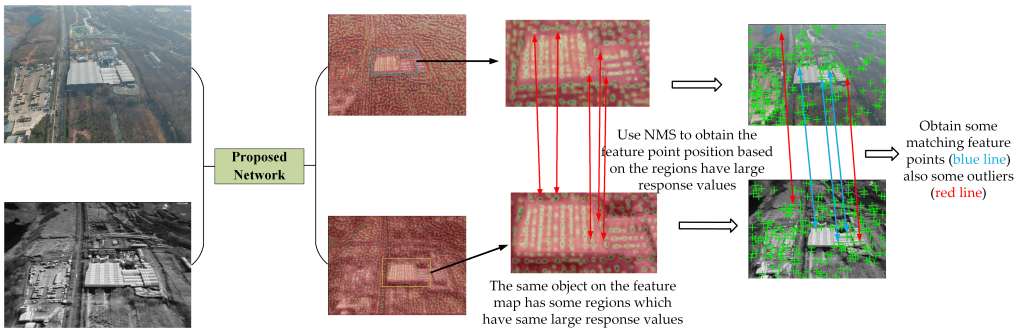

Figure 6. For one given input image pair, we show the pipeline of the valid local features obtained procedure.

3.2. Training Loss

Detector Loss: Existing research [28,30] highlighted that the repeatability of the interest point is a key issue of great importance; however, it cannot be addressed by standard supervised methods. Given the two images with correspondences, every pixel in the first image I has only one correspondence pixel in the second image I'. Thus, a ground truth transformation between the two images for the same scene needs to be obtained. Let $T \epsilon R^{3\times3}$ represent the transformation. Now, T can be estimated using multiple-view geometry [30] or optical flow [31]. Even if only image I is known, the correspondence can be obtained by autogenerating a known transformation (e.g., homography), as performed in the Super-Point [28]. Let F and F' represent feature maps for images I and I', respectively. Then, F'_T is warped from F' based on T. The local maxima in F should correspond to the ones in F'_T. In this work, cosine similarity measures the correspondence between the two feature maps. When cosine similarity is maximized, the two feature maps are identical (i.e., the exact correspondence is achieved). However, in practice, border and occlusions on

images affect the cosine similarity value and may have side effects on the results. Therefore, the feature maps were segmented into many small patches with several overlaps. Next, the cosine similarity between the patches is computed, and the loss is defined as:

$$L_{det-cos}(I, I', T) = 1 - \frac{1}{|P|} \sum_{p \epsilon P} cosine(F[P], F'_T[P]) \tag{3}$$

It is worth noting that when the value on the F and F'_T are converged to a close constant, this loss function will be minimized trivially. To avoid this and consider that the local maxima on the feature maps are feature detection results, we design a novel loss function which tries to separate the local maxima from the other points as follows:

$$L_{l-max}(I) = 1 - \frac{1}{|P|} \sum_{p \epsilon P} (max_{(i,j) \epsilon p} F^a_{ij} - mean_{(i,j) \epsilon p} F^a_{ij})^2 \tag{4}$$

Note that the algorithm can decide the number of detected features through designing the patch size in these loss functions. The final detector loss consists of these two loss terms.

$$L_{dectector}(I) = L_{det-cos}(I, I', T) + \mu(L_{l-max}(I) + L_{l-max}(I')) \tag{5}$$

Descriptor Loss: Our goal is to train the descriptor to find the local features in patches with distinctiveness for feature matching. As in previous works [40–43], descriptor matching is cast as a rank learning problem. First, each descriptor D_{ij} from the first image I represents patch structure information around the interest point (i, j). Then, use this D_{ij} to search the most similar $D'_u v$ in the corresponding image I'. Most of the previous works considered pairwise or tuple-wise loss functions, which use a limited number of patches on the corresponding image I' (Figure 7a). In contrast, our proposed approach uses the knowledge on the ground truth correspondence between images I and I'. More precisely, the descriptor D_{ij} can be compared to the descriptors $\{D'_{uv}\}$ around the corresponding point (u, v) in image I' using the designed listwise losses, as seen in Figure 7b.

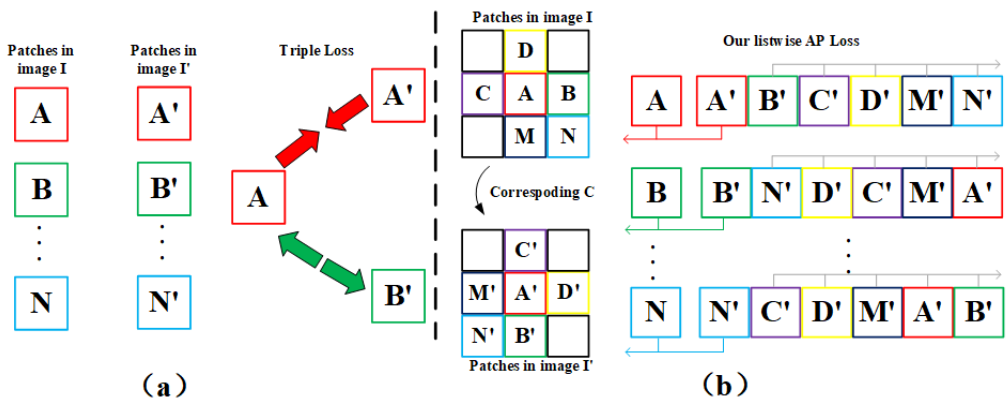

Figure 7. Comparison of descriptor loss. (**a**) is the description of traditional triple loss and (**b**) is the description of our listwise AP loss.

As the most popular loss function for descriptor training, the triple loss intuitively seeks to minimize the distance of the corresponding descriptor d_A from d'_A, while maximizing the distance to other descriptors d'_B or d'_N . However, it only performs local optimization based on paired or ternary patches. The other patches are not involved in the optimization and cannot correct the global metric. In contrast, our proposed descriptor average precision (des-AP) loss considers the patches around the interest point simultaneously and directly optimizes the AP from these patches. Inspired by previous work [43,44], we defined a region of size S around the interest point (i, j) on image I, then compared all the patches around the corresponding point (u, v) on image I'. Let $Q = \{A, B, \ldots, N\}$ denote a batch of

patches around the interest point (i, j) and $Q' = \{A', B', \ldots, N'\}$ denote a batch of patches around the correspondence point (u, v). Using the known ground truth correspondence $C \epsilon R^{H \times W \times 2}$, we can compute the AP for all the patches in the region around the interest point (i, j) and the correspondence point (u, v). These image patches were sorted according to their similarities in decreasing order. The cosine similarity is used to represent the patches similarity:

$$S_i^q = sim(Q_i, Q_j') \tag{6}$$

We need to compute the similarity of each patch in batch Q with the patches in batch Q', which can be represented as:

$$S_{Q-Q'}(i) = \sum_{j=1}^{Q} S_i^q = \sum_{j=1}^{Q} sim(Q_i, Q_j') \tag{7}$$

The training goal is maximizing the AP which is computed by averaging the similarity between patches in batch Q and and patches in batch Q':

$$AP_{Q-Q'}(i) = \frac{1}{Q} \sum_{i=1}^{Q} S_{Q-Q'}(i) \tag{8}$$

So that the descriptor loss for each interest point will be computed as:

$$l_{des-AP}(i) = 1 - AP_{Q-Q'}(i) \tag{9}$$

3.3. Implementation Details

3.3.1. Train Scheme

According to the introduction above, the infrared image and RGB image relies on a few datasets and is faced with the difficulty of obtaining the ground truth value by human annotation. Thus, an iterative learning scheme was designed in this paper. In this training scheme, it embedded the learning-based local features in the most popular open-source structure-from-motion (SFM) tool, COLMAP. COLMAP builds on the multiple-view geometry theory to provide a dense 3D reconstruction and estimate the internal and external camera parameters for overlapping images. For learning-based image matching methods, the most crucial step is obtaining a reliable correspondence for algorithm training with limited infrared aerial image and RGB satellite reference image datasets. Considering these aspects, we use the geometric constraints involved in the COLMAP reconstruction step to compute a more reliable correspondence between RGB-IR image pairs and iteratively optimize the correspondence by reusing the dataset with a trained model.

In the proposed training scheme, the feature detection and matching step were modified to incorporate our premature detector and descriptor, instead of the SIFT in COLMAP. Moreover, epipolar constraints were added to COLMAP to obtain high-quality pseudo-ground truth correspondence for the iterative training of the algorithm. In every iteration, COLMAP and epipolar constraints reject the incorrect matching features before we compute the correspondences of each RGB-IR image pair, while the reliable correspondences are kept as the pseudo-ground correspondences for the training of the algorithm.

In the first step, the visible light images are used to pretrain a model that extracts high-level image features. Alternatively, other state-of-the-art algorithms can be utilized. Then, either the pretrained model or the alternative algorithms are used to replace the feature extraction and matching step in the original COLMAP. Through sparse reconstruction with epipolar constraints and the dense reconstruction, a dense correspondence between the infrared aerial images and satellite reference images is obtained and stored as a fundamental matrix. The fundamental matrix represents each feature's projected location from a visible light image to an infrared image. Finally, the matrix is used in network training to obtain the detector and descriptor for infrared aerial images and satellite reference images. In the first training iteration, the epipolar constraints and outliers were considered. Rejection in

COLMAP helps to reduce the errors in the fundamental matrix. Such errors could have a significant impact on the scene matching results. In order to enable self-evolution of our algorithm, after every training scheme, the trained model replaces the feature extraction and matching step in COLMAP to obtain a more accurate fundamental matrix and consequently results in continuous improvement of the algorithm. The training pipeline is illustrated in Figure 8.

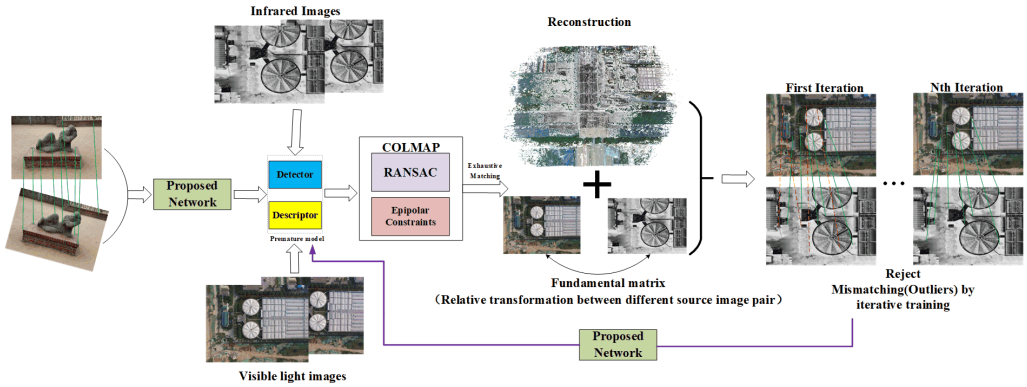

Figure 8. The proposed training pipeline.

Algorithm 1 Train stage 1: Self-labelled local feature detector and descriptor model trained on visible light images.

Input: The visible image I, autogenerated homography matrix: H_i;
Parameters: Interest points function on images: $f(\cdot)$ so that $p = f(I)$ and $p' = f(I')$; Numbers of generated homography matrix: n;
Output: Premature local feature detector and descriptor.

1: Wrapped image I by autogenerated homography matrix H_i to generate image I' $I' = H_i(I)$
2: **for** $i = 1$ to n **do**
3: Used the proposed convolutional neural network to detect the local features (interest points) on image pairs (I and I') with known correspondence. Thus, every interest point: $p = H_i^{-1}P' = H_i^{-1}f(I') = H_i^{-1}f(H_i(I))$
4: Aggregated the detect results to obtain the pseudo-ground truth interest points as the training labels.
5: Repeated training iterations.
6: **end for**
7: Started the joint training with input image I, wrapped image I_i', and the labels obtained by step 4.
8: Computed the detector loss in Formula (5) and the descriptor loss in Formula (8).
9: Until the sum of detector loss and descriptor loss convergence.
10: **return** the premature local feature extraction model

Irrespective of whether the images are rendered from the same or different visual sensors, for a known image pair correspondence, every interest point p in image I has a corresponding point p' in image I'. In this work, p' is a self-labeling result. Therefore, the correspondence c for each image pair is critical for algorithm training. In the training pipeline, the autogenerated homography matrix (H) and the fundamental matrix (F) from COLMAP were used to build the correspondence in different training stages:

$$pH = p' \text{or } pF = p' \tag{10}$$

while the autogenerated homography matrix can be calculated, the fundamental matrix from COLMAP is impacted by mismatching points (i.e., outliers). Therefore, the iterative training stage serves to reject the outliers and obtain a more accurate fundamental matrix

using RANSAC and epipolar constraints. This process improves self-labeling results for algorithm training.

The image matching based on the learning-based features extraction can be expressed in two train stage. The train stage 1 is designed as Algorithm 1 shows.

After the network has been trained on the visible light images, the algorithm already has the ability to find the common local features on the same source image pairs with different illumination and viewpoint conditions and even has the initial ability to find the common features on the visible light and infrared image pairs. However, the proposed network still requires infrared aerial images and satellite reference images with correspondence information to improve its performance. In other words, the method still needs the ability to robustly find common features in the extreme appearance change region. Otherwise, the local features on an RGB-IR image pair found by the algorithm will give rise to a large number of wrong matchings. It will introduce errors when computing the correspondence between the visible target image and the infrared search image. Therefore, we designed the train stage 2 as Algorithm 2 shows.

Algorithm 2 Train stage 2: Self-labelled local feature detector and descriptor model trained on visible light image and infrared image pairs.

Input: The visible light image I_V and the infrared image I_{In}, captured on the same scenes;
Parameters: Fundamental matrix as F; every interest point in I_V as p_V and its correspondence point I_{In} as p_{In}
Output: The local feature detector and descriptor model for RGB-IR image pairs.

Modified the feature extraction method in COLMAP: Used the pretrained detector and descriptor model instead of the original feature extraction method (SIFT).
2: Repeated training scheme.
Used the modified COLMAP to extract local features on image pairs (I_V and I_{In}) then matching the common local features.
4: Based on the result in step 3, obtained the sparse reconstruction result then used the outlier rejection and epipolar constraints to exclude the mismatching local features.
Used the modified COLMAP to generate the dense reconstruction to obtain the F, so that $p_V = F^{-1} p_I n$. Use this correspondence as the training labels.
6: Repeated training iterations.
Started the joint training with image I_V and I_{In} captured on the same scenes and the labels obtained by step 5.
8: Computed the detector loss in Formula (5) and the descriptor loss in Formula (8).
Until the sum of detector loss and descriptor loss convergence.
10: Until the modified COLMAP matching result has no wrong.
return the final local feature extraction model for visible light image and infrared image pairs

Every training scheme iteration will generate a model for local features detection and description. Using this trained model, the common local features between visible light and infrared images will be found for computing the correspondence relation (homography matrix) of the visible light reference image and infrared real-time image. Then the visible light reference image is precisely matched on the infrared real-time image.

3.3.2. Training Data

Two types of data were used to train the proposed method: (a) Popular image retrieval and 3D reconstruction datasets (MS-COCO [45] and MegaDepth [46]), which includes only visible light images (more than 30,000 images). These datasets were used to train the premature model. (b) The reference and real-time image pair dataset. The reference images were collected from the satellite (we downloaded from Google Earth) and DJI drone with the visible light camera. The real-time images were collected by a DJI drone with the infrared camera (including 8000 image pairs and all pairs of images with at least 50% overlap). These images were used to train the final model. These images used for

model training are not included in the test dataset for the comparison experiment. Table 2 summarizes the detail of the training and validation data. Figure 9 shows the data samples of different training stages.

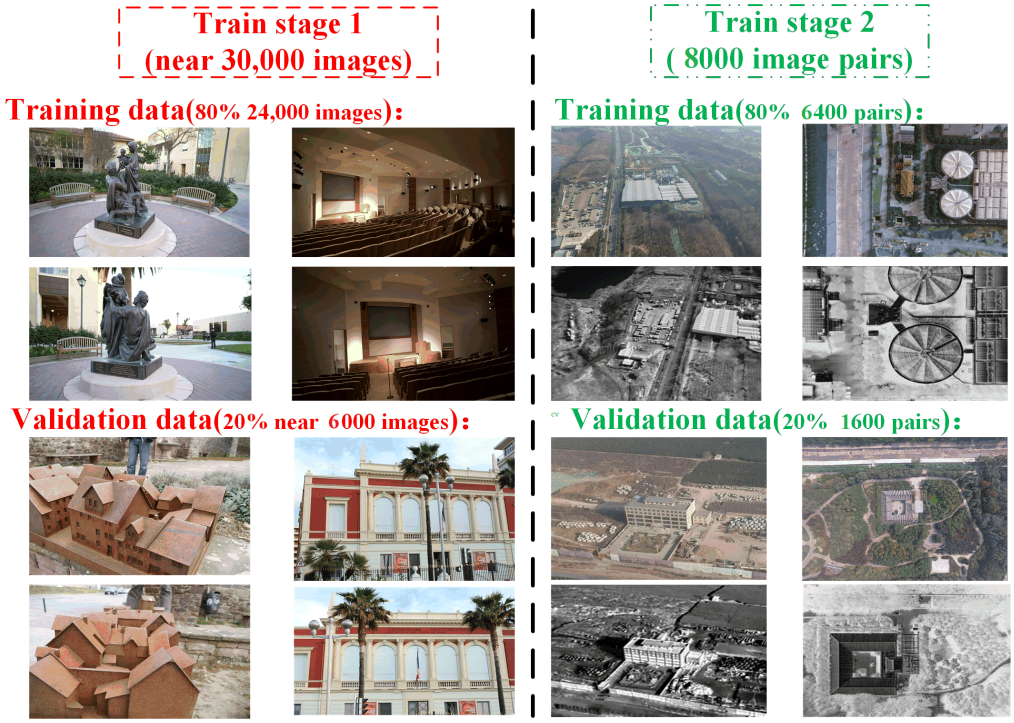

Figure 9. The proposed training pipeline.

Table 2. The detail of the training and validation data.

Dataset for Train Stage	Source	Number of Images	Resolution
Train Stage 1	MS-COCO [45]	near 30,000 images 80% for training	640 × 480
	MegaDepth [46]	20% for training	640 × 480
Train Stage 2	RGB images from satellite	8000 image pairs	720 × 512
	RGB images from UAV	80% for training	720 × 512
	infrared images from UAV	20% for training	720 × 512

3.3.3. Model Training

The design, training, and evaluation of the learning local feature extraction model were embedded in the PyTorch framework. The training dataset was randomized by shuffling and was fed into batches of size 64 [47]. The model was trained for up to 100 epochs on the training datasets. For the model design and fitting, we adopted an Adam optimizer and used the step-wise learning rate decay strategy with an initial learning rate of 0.0005, which decayed at a rate of 0.95 after every 20 epochs to make our model coverage faster with a higher accuracy [48–51]. To accelerate the training, we used four NVIDIA GTX TITAN X GPUs and a multi-GPU training mechanism in PyTorch. All of the hyperparameters inside the networks were identical for a fair comparison.

4. Experiment and Results

In this section, we first introduce the datasets used in the experiments and the details of experimental implementation. We then display image matching results on the infrared aerial (UAV captured) images and satellite reference images and provide qualitative and quantitative comparisons between our proposed method and the other comparison meth-

ods. In the end, we test the precision and efficiency of all the comparison methods through localization application of UAV in the real scenario.

4.1. Experimental Data and Metrics

Datasets: The developed method was evaluated on the infrared and visible light images dataset, which was captured by the DJI drone in the Xi'an urban area and satellite images from Google Earth (Figure 10 shows the sample of these images). Experiments using the dataset serve to demonstrate that our method can obtain the state-of-the-art performance for the infrared aerial (UAV captured) image and satellite reference image matching task. The dataset included 20 different scenes, and each scene had 100 pairs of infrared (in size 720 × 512) and visible light (RGB) images (in size 1280 × 960; we resized the RGB images to 640 × 480 before sending them into the network).

Figure 10. RGB-IR image pairs from our test dataset. (**a**) is a visible reference image with the target scene obtained from Google Earth. (**b**,**c**) are the infrared images with the target scene captured at different viewpoint.

Metrics: This work uses well-known and widely used evaluation metrics: precision, recall, and F1-score. Each metric is briefly described as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{13}$$

Precision denotes the quotient of correctly detected targets when the target is present in infrared images (true positive, TP) and the sum of TP and the number of cases where a target was detected despite the target not being present in infrared images (false positive, FP). Recall represents the quotient of TP and the sum of TP and the number of cases where a target was not detected despite the target being in infrared images (false negative, FN). F1-score is a trade-off between recall and precision. The target regions define the calculation of the TP, FP, TN, and FN using the intersection over union (IoU) as follows:

$$IoU = \frac{DetectionRestult \cap GroundTruth}{DetectionRestult \cup GroundTruth} \tag{14}$$

$$\begin{cases} \text{if target region in infrared image and IoU} > 0.6 & TP + 1 \\ \text{else} & FN + 1 \\ \text{if no target region in infrared image and IoU} < 0.1 & TN + 1 \\ \text{else} & FP + 1 \end{cases} \tag{15}$$

4.2. Experimental Results of Comparison Methods

The proposed method was contrasted with the state-of-the-art methods for scene matching, which include the handcrafted method SIFT [14] and learning-based methods, including Super-Point [28], D2-Net [30], and R2D2 [31]. All of the learning-based local

feature extraction methods are re-trained or fine-tuned on the training dataset before the comparison experiments. In the experiments, the target scene from reference visible light images was selected. Then, local features were extracted and the descriptors were computed to match the reference visible light images with the real-time infrared images. The RANSAC method implemented in OpenCV was utilized to reject the outliers, while the inliers were used to compute the homography matrix for target scene location identification on the infrared image.

Experiment on RGB-IR Image Pairs: First, the proposed method is compared to the state-of-the-art methods on RGB-IR image datasets. Tables 3 and 4 show the results of the corresponding experiments.

Table 3. RGB-IR image pairs matching results using 6K features. Precision (%), recall (%), and F1-score at different viewpoint change thresholds are reported.

Viewpoint Change (deg)	**0/30/45**	**0/30/45**	**0/30/45**
Method	**Precision (%)**	**Recall (%)**	**F1-Score**
SIFT [14]	33.2/19.6/8.7	10.1/9.9/5.3	15.5/13.2/6.6
DoG-HardNet [36]	71.2/65.1/48.7	65.2/61.6/41.7	68.1/63.3/44.9
Super-Point [28]	69.1/64.5/43.9	59.8/55.3/40.1	64.1/59.5/41.9
D2-Net [30]	78.4/56.3/40.3	68.7/56.7/37.8	73.2/56.5/39.0
R2D2 [31]	77.6/65.4/45.6	67.3/62.5/42.4	72.1/63.9/43.9
Our Method (1st train Iteration)	**78.5/64.1/49.3**	**68.9/60.8/46.3**	**73.4/62.4/47.8**
Our Method (5th train Iteration)	**81.2/66.3/50.4**	**72.3/63.1/48.7**	**76.5/64.7/49.5**

Table 4. RGB-IR image pairs matching results using 2K features. Precision (%), recall (%), and F1-score at different viewpoint change thresholds are reported.

Viewpoint Change (deg)	**0/30/45**	**0/30/45**	**0/30/45**
Method	**Precision (%)**	**Recall (%)**	**F1-Score**
SIFT [14]	26.8/16.6/5.7	8.7/4.5/2.2	13.1/7.0/3.2
DoG-HardNet [36]	50.5/41.4/34.7	43.7/39.8/32.6	46.9/40.6/33.6
Super-Point [28]	50.3/41.1/33.9	40.0/37.1/29.4	44.6/39.0/31.5
D2-Net [30]	52.5/39.0/32.2	46.0/38.2/27.0	49.0/38.6/29.8
R2D2 [31]	51.2/40.9/33.8	45.1/39.2/30.6	48.0/40.0/32.1
Our Method (1st train Iteration)	**53.7/42.5/35.6**	**48.3/41.9/33.7**	**50.9/42.2/34.6**
Our Method (5th train Iteration)	**56.4/42.7/36.9**	**50.4/42.3/35.2**	**53.2/42.5/36.0**

To visually assess the effect of our method and the others, in Figure 11, we compare these methods by recognizing the target image (Figure 11a visible light image) on the search image (Figure 11b infrared image), with 30 degrees of viewpoint change. The scene matching results of all methods are shown in Figure 11c–j, which shows the comparison results of the IoU. The results of SIFT show (Figure 11c) that the handcrafted feature-based methods can hardly find the common features between infrared aerial images and satellite reference images to matching and recognizing the target region on the satellite reference image. The results shown in Figure 11e–i proved that the learning feature-based method could employ the local features to recognize the target in a close region but is not accurate enough. Our method (Figure 11h) is better and more accurate than the other compared methods. Moreover, the iterative training scheme for our method can significantly improve the infrared aerial images and satellite reference matching results, as shown in Figure 11h,i.

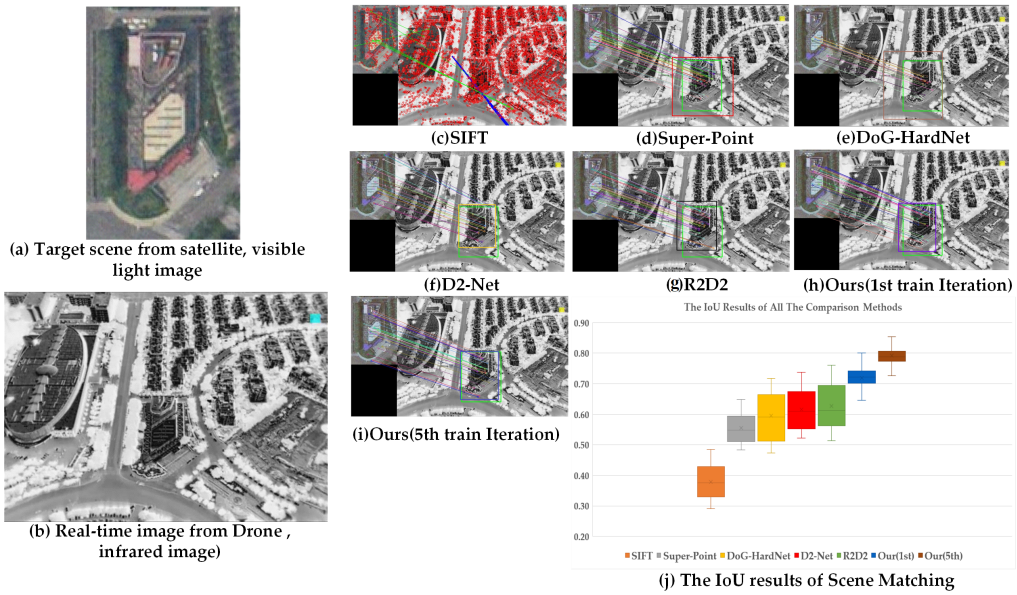

Figure 11. RGB-IR image pairs sample from our test dataset and the image matching results of all the comparison methods. (**a**) is a reference visible light image from Google Earth. (**b**) is the real-time infrared image captured at different view-point. (**c**) is the matching result of sample image from SIFT. (**d**) is the matching result of sample image from Super-Point. (**e**) is the matching result of sample image from DoG-HardNet. (**f**) is the matching result of sample image from D2-Net. (**g**) is the matching result of sample image from R2D2. (**h**) is the matching result of sample image from the first train iteration of our method. (**i**) is the matching result of sample image from the fifth train iteration of our method. (**j**) is the IoU results of all the image matching tests.

4.3. Applied Experiment

Finally, we apply all of the comparison methods to trajectory reconstruction for UAV navigation via matching real-time images on the satellite image. The real-time images were captured by infrared camera on UAV and the satellite image was download from Google Earth. The real-time images (720 × 510) are shown in Figure 12. The satellite images (8000 × 6000 with 2 m resolution) are shown in Figure 13. As shown in these figures, the extreme image appearance changes from different imaging sensor and large geometric change caused by different capture time and different platform can be intuitively found.

Figure 12. Samples of the real-time infrared images.

Figure 13. Reference satellite image download from Google Earth.

To achieve fast image matching, we cut the reference image into several subimages. Several image matching results are shown in Figure 14. The result shows that our proposed method is able to deal with the extreme image appearance changes and large geometric changes.

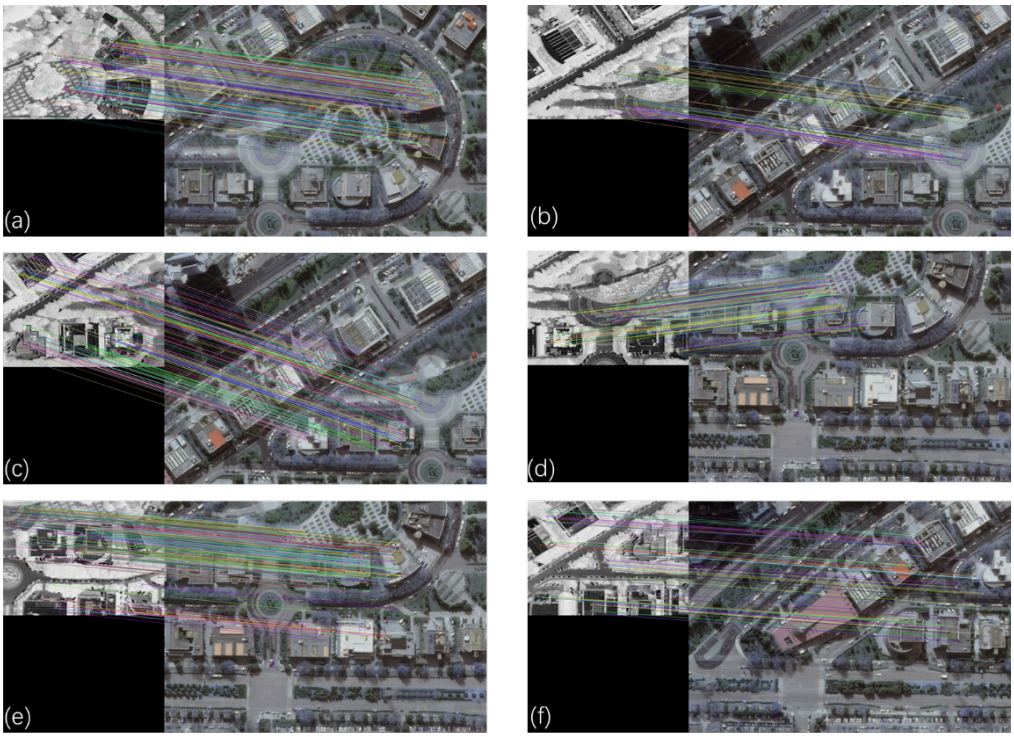

Figure 14. Six image matching (**a**–**f**) results of the samples in Figure 12 and the subimages of the reference satellite image in Figure 13.

To assess the localization precision, root-mean-square error (RMSE) is used in the overlapped image region between the real-time infrared images and the reference satellite image, which is calculated as follows:

$$RMSE = \frac{\sqrt{(x_i^{ref-sat} - x_i^{rt-inf})^2 + (y_i^{ref-sat} - y_i^{rt-inf})^2}}{N}, i = 1, \ldots, N \tag{16}$$

where $(x_i^{ref-sat}, y_i^{ref-sat})$, $(x_i^{rt-inf}, y_i^{rt-inf})$ are the corresponding pixels we selected in the overlapped area from the reference image and the transformed infrared images, respectively. Each pixel on the reference satellite image has corresponding latitude and longitude in the World Coordinate System. N is the number of corresponding pixels we selected. Thus, we can also assess the localization error in physical distance (in meters), which is calculated as (17):

$$Distance = 2arcsin\sqrt{\frac{sin^2(Lat1 - Lat2)}{2} + cos(Lat1) \times cos(Lat1) \times \frac{sin^2(Lng1 - Lng2)}{2}} \times 6378.137 \tag{17}$$

where $(Lat1, Lng1)$ is the corresponding latitude and longitude for $(x_i^{ref-sat}, y_i^{ref-sat})$ and $(Lat2, Lng2)$ is the corresponding latitude and longitude for $(x_i^{rt-inf}, y_i^{rt-inf})$; the 6378.137 (in kilometers) is the radius of earth. The RMSE and the localization error (LE) results of the samples in Figure 10 are shown in Table 3.

Figure 15 shows the UAV trajectory tracking results of our method, and the displayed results correspond to the results shown in Figure 14.

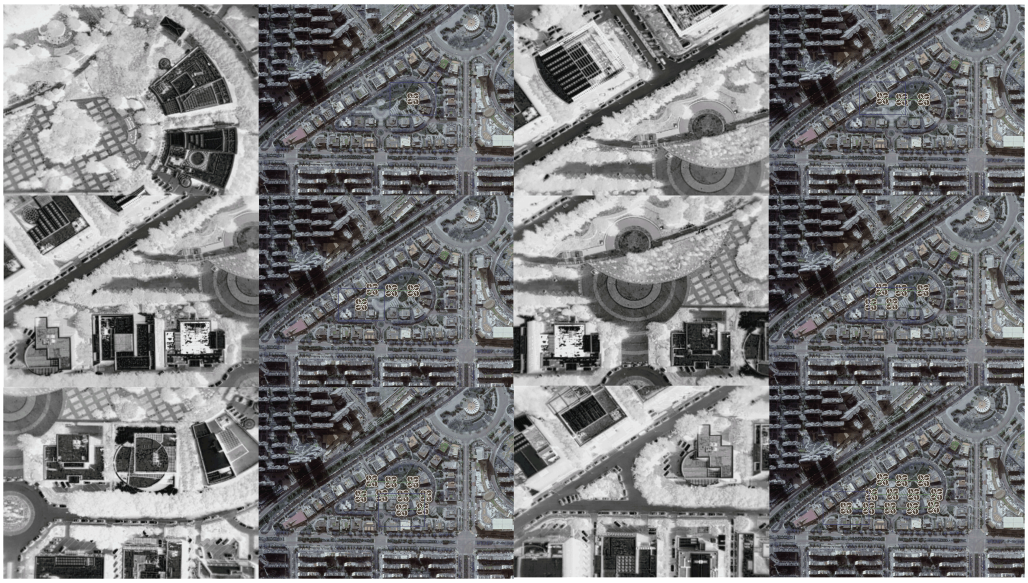

Figure 15. The UAV trajectory tracking results of our method.

5. Discussion

5.1. Discussion on Experimental Results of the Infrared Aerial Images and Satellite Reference Images

Our overall goal is to robustly match the same target region in a subpixel-level accuracy from the infrared aerial image and satellite reference image. The proposed method reaches the goal through two steps. First, the detector branch of proposed LLFE can extract local features at a subpixel level and the descriptor branch can jointly learn structure information from the region around the local features for matching the local features on the infrared aerial image and satellite reference image. Second, we designed the iterative training scheme to automatically produce the pseudo-ground truth correspondence for detector and descriptor training. Meanwhile, the modified COLMAP can provide more

precise outlier rejection to improve the quantity of pseudo-ground truth, which leads to the improved performance of the algorithm. As can be observed in Table 3, the precision and F1-scores of SIFT are lower than other methods, of which the major cause lies in the fact that SIFT is a handcrafted feature that cannot well adapt the extreme appearance changes due to the different imaging principles of visible light and infrared images. Among learning-based local features, the metrics show that the joint learned descriptors and detectors achieve positive performance, which proves that joint feature extraction and description approaches are appropriate for infrared aerial image and satellite reference image matching. However, the existing learning-based methods do not establish a corresponding relationship between infrared images and RGB images during training. Therefore, even using the fine-tune on the infrared and RGB images for these methods, they only focus on feature extraction and descriptor computation on a single image. As the results show in Table 3, the performance of our method, especially the results after five times iterative training, is 2–5% higher than the best performance in the other methods. This is owing to the reason that we utilize the more reliable local feature structure for feature description and take all the advantage of the limited image data by using geometric constraint to find the corresponding relation between the RGB-IR image pairs. In Table 5, we limited the number of extraction local features, and the metrics show that joint feature extraction and description approaches have good robustness even if the number of matching features is significantly reduced. In other words, the learning-based local features are more reliable for infrared aerial image and satellite reference image matching. The result of our method shows the smallest performance reduction which proved that our method provides a reliable feature descriptor that brings robustness to the local feature matching. At the same time, the iterative training scheme can provide performance improvement even if the number of extraction local features are limited. It proves that the geometric constraint can help to generate the high quality pseudo-ground correspondences which will lead the network to find the robust local feature for image matching by the iterative training scheme. Compared with the literature, the results demonstrate that the proposed algorithm significantly outperforms the state-of-the-art image matching based on local feature extraction. Furthermore, there is sufficient evidence to conclude that the developed algorithm represents greater robustness than other methods for images with viewpoint change. Moreover, the iterative train scheme is able to improve the performance of the algorithm through providing more reliable and accurate correspondence for algorithm training, even with the limited amount of training datasets.

Table 5. The RMSE (in pixel) and localization error (in meters) results on the six sample image pairs. The SIFT and deep compare cannot finish this experiment; thus, their results are not displayed here.

Samples in Figure 10	SIFT [14] ($RMSE/LE$)	DoG-Hard [36] ($RMSE/LE$)	Super-Point [28] ($RMSE/LE$)	D2-Net [30] ($RMSE/LE$)	R2D2 [31] ($RMSE/LE$)	Our Method
(a)	—/—	—/—	—/—	3.40/7.16	2.14/4.32	**1.69/3.61**
(b)	2.11/4.29	3.31/7.09	4.51/8.69	5.14/10.23	2.35/4.81	**2.04/4.15**
(c)	1.76/3.17	2.76/5.71	2.78/5.63	3.51/7.93	1.98/3.88	**0.82/1.92**
(d)	—/—	—/—	—/—	6.71/13.43	5.62/11.27	**3.97/7.78**
(e)	—/—	4.18/8.63	5.27/12.11	5.73/12.13	3.53/7.12	**2.44/5.03**
(f)	—/—	4.74/9.66	6.25/12.98	6.31/13.03	4.84/9.25	**3.71/7.34**

5.2. Discussion on Applied Experimental Results

As demonstrated by the RMSE and localization error in Table 5, the localization precision of the DOG-HardNet [36] can reach subpixel level when enough matching features are detected, as the DOG-HardNet has the same feature detection principle as SIFT [14]. The Super-Point [28], D2-Net [30], and R2D2 [31] only have pixel-level feature detection precision; therefore, even though they matched enough local features, their localization precisions are lower than DOG-HardNet and our method. On the other hand, the extracted local features from these comparison methods still have some outliers that also

impact the accuracy of the transformation model computing between the infrared aerial image and the satellite reference image. Therefore, the large error will occur for the UAV localization application. The main reason for our method's superiority is that we modified the feature detector branch which can provide the subpixel-level local feature detection. Meanwhile, the local features we detected are verified by the geometric constraint, which means the detected local features do not contain outliers. The joint network architecture also makes our method reduce the computing time cost. In our computer with Intel Xeon E5-2637MQ CPU at 3.50 GHz, the aerial infrared image and satellite RGB image matching time can be shortened to 2 s when we set the number of features as 2000. Therefore, our proposed method has the potential localization ability for UAV navigation application.

6. Conclusions

This paper has proposed a novel approach for infrared aerial image and satellite reference image matching through local feature extraction. Firstly, we designed a joint network for local feature detection and description. The feature detector branch provides subpixel feature localization and the feature descriptor gives the structure information of the patch around the feature point to enhance the distinctiveness of local features. The proposed learning local feature combines these advantages, which can obtain more accuracy and robustness correspondence in RGB-IR image pairs for the image matching task. Secondly, in attempting to solve the problem of scarce training data caused by the limited amount of infrared aerial images and satellite reference images, we propose an iterative train scheme which significantly improves the performance of our algorithm by self-labeling the reliable pseudo-ground truth correspondence. To evaluate the performance of our method, we conducted image matching experiments with the real-time infrared images captured by the UAV and the reference images captured by the satellite, including 20 different image samples and 2000 images in total. The experimental results have shown our method surpasses the state-of-the-art methods of infrared aerial image and satellite reference image matching with a different viewpoint change condition. The F1-score of the matching results further increases by a large margin (up to 4–6%) compared to the other methods. In addition, we demonstrated the effectiveness of our method in the application of UAV localization by matching the target regions on the real-time infrared images captured by UAV with the reference satellite images. The RMSE of the proposed method surpasses the state-of-the-art method by up to nearly 25%. Meanwhile, the localization error results of the proposed method decrease by 20% compared to the state-of-the-art method.

As the image matching approach is based on local feature extraction and matching, we used the learning-based local feature extraction with the traditional matching algorithm in this paper. However, the learning-based matching algorithms have recently outperformed traditional methods. Therefore, future work can be carried out in the learning-based matching algorithm that can provide an appropriate approach to find the correspondence between infrared aerial image and satellite reference image matching. On the other hand, for the real-time application of the proposed method, a lot of work is still needed, such as model pruning and quantization, for the model compression to improve the operating time of the proposed method to reach the real-time application requirement.

Author Contributions: Methodology, X.Z. and Z.H.; resources, Z.M.; software, X.Z.; writing—review and editing, X.Z., Z.M., Z.W. and L.W. All authors have read and agreed to the published version of the manuscript.

Funding: The Qian Xuesen Youth Innovation Foundation from the China Aerospace Science and Technology Corporation (grant number 2019JY39).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the Ninth Academy of the China Aerospace Science and Technology Corporation for supporting this research.

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

References

1. Wang, C.; Peng, T.; Hu, L.; Liu, G. Improved UAV scene matching algorithm based on censure features and FREAK descriptor. In *International Conference on Computer Engineering and Networks*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 158–167.
2. Kaniewski, P.; Grzywacz, W. Visual-based navigation system for unmanned aerial vehicles. In Proceedings of the 2017 Signal Processing Symposium (SPSympo), Jachranka Village, Poland, 12–14 September 2017; pp. 1–6. [CrossRef]
3. Zhuo, X.; Koch, T.; Kurz, F.; Fraundorfer, F.; Reinartz, P. Automatic UAV image geo-registration by matching UAV images to georeferenced image data. *Remote Sens.* **2017**, *9*, 376. [CrossRef]
4. Ebadi, K.; Wood, S. Scene matching-based localization of unmanned aerial vehicles in unstructured environments. In Proceedings of the IEEE 2018 52nd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 28–31 October 2018; pp. 1519–1523.
5. Liu, J.S.; Liu, H.C. Visual Navigation for UAVs Landing on Accessory Building Floor. In Proceedings of the IEEE 2020 International Conference on Pervasive Artificial Intelligence (ICPAI), Taipei, Taiwan, 3–5 December 2020; pp. 158–163.
6. Andrew, A.M.; Hartley, R. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2001.
7. Conte, G.; Doherty, P. An integrated UAV navigation system based on aerial image matching. In Proceedings of the 2008 IEEE Aerospace Conference, Big Sky, MT, USA, 1–8 March 2008; pp. 1–10.
8. Balamurugan, G.; Valarmathi, J.; Naidu, V. Survey on UAV navigation in GPS denied environments. In Proceedings of the IEEE 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Odisha, India, 3–5 October 2016; pp. 198–204.
9. Wan, X.; Liu, J.; Yan, H.; Morgan, G.L. Illumination-invariant image matching for autonomous UAV localisation based on optical sensing. *ISPRS J. Photogramm. Remote Sens.* **2016**, *119*, 198–213. [CrossRef]
10. Carr, J.R.; Sobek, J.S. Digital scene matching area correlator (DSMAC). In *Image Processing For Missile Guidance*; International Society for Optics and Photonics: Bellingham, WA, USA, 1980; Volume 238, pp. 36–41.
11. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the IEEE 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
12. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 404–417.
13. Bellavia, F.; Colombo, C. Is there anything new to say about SIFT matching? *Int. J. Comput. Vis.* **2020**, *128*, 1–20. [CrossRef]
14. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
15. Held, D.; Thrun, S.; Savarese, S. Deep learning for single-view instance recognition. *arXiv* **2015**, arXiv:1507.08286.
16. Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; Berg, A.C. Matchnet: Unifying feature and metric learning for patch-based matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3279–3286.
17. Zagoruyko, S.; Komodakis, N. Deep compare: A study on using convolutional neural networks to compare image patches. *Comput. Vis. Image Underst.* **2017**, *164*, 38–55. [CrossRef]
18. Tian, Y.; Fan, B.; Wu, F. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 661–669.
19. Kern, J.P.; Pattichis, M.S. Robust multispectral image registration using mutual-information models. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45*, 1494–1505. [CrossRef]
20. Lian-Fa, B.; Jing, H.; Yi, Z.; Qian, C. Registration algorithm of infrared and visible images based on improved gradient normalized mutual information and particle swarm optimization. *Infrared Laser Eng.* **2012**, *41*, 248–254.
21. Torabi, A.; Bilodeau, G.A. Local self-similarity-based registration of human ROIs in pairs of stereo thermal-visible videos. *Pattern Recognit.* **2013**, *46*, 578–589. [CrossRef]
22. Bhuiyan, M.A.E.; Witharana, C.; Liljedahl, A.K. Use of Very High Spatial Resolution Commercial Satellite Imagery and Deep Learning to Automatically Map Ice-Wedge Polygons across Tundra Vegetation Types. *J. Imaging* **2020**, *6*, 137. [CrossRef]
23. Zhang, W.; Liljedahl, A.K.; Kanevskiy, M.; Epstein, H.E.; Jones, B.M.; Jorgenson, M.T.; Kent, K. Transferability of the deep learning mask R-CNN model for automated mapping of ice-wedge polygons in high-resolution satellite and UAV images. *Remote Sens.* **2020**, *12*, 1085. [CrossRef]
24. Szeliski, R. *Computer Vision: Algorithms and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2010.
25. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
26. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 467–483.
27. Noh, H.; Araujo, A.; Sim, J.; Weyand, T.; Han, B. Large-scale image retrieval with attentive deep local features. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3456–3465.
28. DeTone, D.; Malisiewicz, T.; Rabinovich, A. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 224–236.

29. Ono, Y.; Trulls, E.; Fua, P.; Yi, K.M. LF-Net: Learning local features from images. *arXiv* **2018**, arXiv:1805.09662.
30. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-net: A trainable cnn for joint detection and description of local features. *arXiv* **2019**, arXiv:1905.03561.
31. Revaud, J.; Weinzaepfel, P.; De Souza, C.; Pion, N.; Csurka, G.; Cabon, Y.; Humenberger, M. R2D2: Repeatable and reliable detector and descriptor. *arXiv* **2019**, arXiv:1906.06195.
32. Savinov, N.; Seki, A.; Ladicky, L.; Sattler, T.; Pollefeys, M. Quad-networks: Unsupervised learning to rank for interest point detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1822–1830.
33. Zhang, L.; Rusinkiewicz, S. Learning to detect features in texture images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6325–6333.
34. Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; Moreno-Noguer, F. Discriminative learning of deep convolutional feature point descriptors. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 118–126.
35. Simonyan, K.; Vedaldi, A.; Zisserman, A. Learning local feature descriptors using convex optimisation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1573–1585. [CrossRef]
36. Mishchuk, A.; Mishkin, D.; Radenovic, F.; Matas, J. Working hard to know your neighbor's margins: Local descriptor learning loss. *arXiv* **2017**, arXiv:1705.10872.
37. Schonberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.
38. Fisher, A.; Cannizzaro, R.; Cochrane, M.; Nagahawatte, C.; Palmer, J.L. ColMap: A memory-efficient occupancy grid mapping framework. *Robot. Auton. Syst.* **2021**, *142*, 103755. [CrossRef]
39. Ma, J.; Jiang, X.; Fan, A.; Jiang, J.; Yan, J. Image matching from handcrafted to deep features: A survey. *Int. J. Comput. Vis.* **2021**, *129*, 23–79. [CrossRef]
40. Revaud, J.; Almazán, J.; Rezende, R.S.; Souza, C.R.D. Learning with average precision: Training image retrieval with a listwise loss. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 5107–5116.
41. Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; Hullender, G. Learning to rank using gradient descent. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; pp. 89–96.
42. Cao, Z.; Qin, T.; Liu, T.Y.; Tsai, M.F.; Li, H. Learning to rank: From pairwise approach to listwise approach. In Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 129–136.
43. He, K.; Lu, Y.; Sclaroff, S. Local descriptors optimized for average precision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 596–605.
44. Cakir, F.; He, K.; Xia, X.; Kulis, B.; Sclaroff, S. Deep metric learning to rank. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1861–1870.
45. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
46. Li, Z.; Snavely, N. Megadepth: Learning single-view depth prediction from internet photos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2041–2050.
47. Radiuk, P.M. Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. *Inf. Technol. Manag. Sci.* **2017**, *20*, 20–24. [CrossRef]
48. You, K.; Long, M.; Wang, J.; Jordan, M.I. How does learning rate decay help modern neural networks? *arXiv* **2019**, arXiv:1908.01878.
49. Ge, R.; Kakade, S.M.; Kidambi, R.; Netrapalli, P. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *arXiv* **2019**, arXiv:1904.12838.
50. Mishra, P. Supervised Learning Using PyTorch. In *PyTorch Recipes*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 127–149.
51. Lewkowycz, A. How to decay your learning rate. *arXiv* **2021**, arXiv:2103.12682.