Hindawi
International Journal of Aerospace Engineering
Volume 2023, Article ID 1135946, 16 pages
https://doi.org/10.1155/2023/1135946

Research Article

PSiamRML: Target Recognition and Matching Integrated Localization Algorithm Based on Pseudo-Siamese Network

Jiwei Fan, Xiaogang Yang, Ruitao Lu, Xueli Xie, and Siyu Wang

PLA Rocket Force University of Engineering, Xi'an 710025, China

Correspondence should be addressed to Xiaogang Yang; doctoryxg@163.com

Received 13 January 2023; Revised 26 August 2023; Accepted 7 September 2023; Published 12 October 2023

Academic Editor: Guillermo Valencia-Palomo

The positioning function of unmanned aerial vehicles (UAVs) is a challenging and fundamental research topic and is the premise for UAVs to realize autonomous navigation. The disappearance of satellite signals makes it challenging to achieve accurate positioning. Thus, visual positioning algorithms based on computer vision have been proposed in recent years and these algorithms have produced good results. However, these algorithms have relatively simple functions and cannot perceive the environment. Their versatility is poor, and mismatching often occurs, which affects the positioning accuracy. Aiming to address the need for integrated target recognition, target matching, and positioning of UAVs, we propose an algorithm that integrates the target recognition, matching, and positioning functions by combining the single-shot multibox detector (SSD) algorithm with the deep feature matching algorithm. This algorithm is based on the idea of pseudo-Siamese networks and the SSD algorithm, introducing a deep feature matching method to directly calculate the correspondence between two images. The main idea is to use the VGG network trained by the SSD target recognition algorithm to extract deep features, without any special training for feature matching. Finally, by sharing neural network weights, the integrated design of target recognition and image-matching localization algorithms is achieved. Mismatches between the real-time and reference images are addressed by introducing the grid-based motion statistics algorithm to optimize the matching result and improve the correct matching efficiency of the target. The University-Release dataset was used to compare and analyze the performance of the proposed algorithm to verify its superiority and feasibility. The results show that the matching accuracy of the PSiamRML algorithm is generally good and that it significantly compensates for changes in the contrast, scale, brightness, blur, deformation, and so on, apart from improving the stability and robustness. Finally, a matching test scenario with aerial images captured by an S1000 six-rotor UAV served to verify the effectiveness and practicability of the PSiamRML algorithm.

1. Introduction

In recent years, the rapid development of cutting-edge technologies such as artificial intelligence and robotics has resulted in the intelligence of unmanned systems becoming a hot research topic in the field of artificial intelligence. The key technology of the new generation of intelligent unmanned systems is based on algorithms and data, focusing on improving the perception, computing, cognitive reasoning, and combat execution capabilities of these systems, thus forming an open, compatible, stable, and mature technology system. Algorithms are at the core of artificial intelligence technology, which forms the core of an unmanned system and provides the basis on which the system performs various actions [1–3]. At present, because various types of aircraft, such as missiles and unmanned aerial vehicles (UAVs), need a navigation system to continuously determine their position to adjust their operating state during the execution of tasks, further research on the UAV navigation system is important. UAVs typically rely on global satellite navigation, inertial navigation, and visual navigation methods for guidance [4–6]. Traditional satellite navigation and positioning methods rely on external satellite signals that are susceptible to environmental and enemy interference. This particularly affects low-altitude UAVs, as satellite signals are easily blocked by high-rise buildings and cannot provide accurate positioning information [7–11]. The inertial navigation system experiences obvious data drift during longer working periods and cannot provide accurate positioning information for extended periods of time. Visual navigation is an autonomous navigation technology that uses image processing, computer vision, and other technologies to acquire motion information and spatial information about UAVs during operation [12]. The low cost of visual navigation, together with its strong anti-interference ability and good positioning effect, has resulted in this navigation technology becoming an important research direction for the autonomous in-flight navigation of UAVs. Visual navigation and positioning technology can be divided into two types according to whether prior information is used as a reference: The first entails using a sequence of UAV aerial images to match and obtain the pose transformation relationship. Mature technologies include simultaneous localization and mapping (SLAM) and visual odometry (VO) [13–15]. The second technology involves the use of a reference image with known coordinate information to complete the localization using image-matching technology [16]. SLAM can complete the localization while building a map using real-time images captured in an unknown environment and is widely used in various indoor localization scenarios. However, it is less effective in open outdoor environments [17]. VO uses a sequence of aerial images to calculate the transformation relationship between the position and attitude of the UAV. The principle on which VO is based is similar to that of the inertial navigation system, and the positioning results also deviate considerably over time [18]. The positioning method that uses reference images for image matching needs to capture ground objects and scenes as reference images in a predetermined matching area in advance and mark real geographic information as a reference image database, which is stored in the onboard computer of the UAV [19]. When the UAV flies to the predetermined area, the real-time images captured by the airborne sensor of the UAV are matched with the reference image in the airborne computer, enabling the current position of the UAV to be determined accurately. Therefore, image-matching navigation is an absolute positioning technology for UAV navigation [20, 21] and guarantees accurate positioning for UAVs undertaking extended flights. Image-matching technology is the process of spatially aligning two images acquired by the same or different sensors of the same area to determine the position relationship between the two images. The main purpose of this technology is to search for the best matching position of the real image in the reference image and provide basic data for the position change of the carrier [22, 23]. Existing image-matching algorithms are mainly divided into grayscale-based and feature-based image-matching methods [24–28]. Grayscale-based image matching takes the template image as a sliding window image, slides through the image in a sequence according to a certain step size, and selects the part with the largest similarity as the final result. Unfortunately, the poor real-time performance and robustness of the gray-scale image-matching method render it unsuitable for the computationally intensive matching task of visual navigation. The feature-based image-matching method overcomes the shortcomings of the gray-scale image-matching method by offering good robustness to various changes, fast calculation, and good matching accuracy. Therefore, researchers in the field of image matching have focused on this method.

2. Related Work

The first problem the UAV has to solve when undertaking a mission is to determine its exact location in the working environment. This means that localization is the basis for UAV perception and decision-making in unknown environments. The accuracy, robustness, and real-time performance of UAV positioning algorithms have a crucial impact on enhancing the autonomous decision-making and combat capabilities of UAVs while improving their overall performance. Feature extraction is an important function of the feature-based image-matching method. Good features should be stable, reliable, repeatable, and moderate in number. In this way, we can ensure that the same features can be extracted from different images of the same scene. Traditional feature extraction methods include scale-invariant feature transform (SIFT) [29], speeded up robust features (SURF) [30], oriented fast and rotated brief (ORB) [31], affine SIFT (ASIFT) [32], binary robust invariant scalable keypoints (BRISK) [33], and binary fisheye spherical distorted robust independent elementary features (FSD-BRIEF) [34]. These algorithms rely on hand-designed feature descriptors; thus, their real-time performance and robustness need to be further improved. Many feature-based matching algorithms that are based on these classical algorithms have since been optimized and improved [35, 36]. The widespread application of convolutional neural networks (CNNs) has led to the proposal of many feature extraction methods based on these networks. The features extracted by deep learning methods have stronger description ability than those extracted by traditional methods and can identify certain features based on the semantic level. Recent research efforts have focused on image matching with the aid of deep learning. Daniel et al. [37] proposed a self-supervised training-based network model for feature point detection and descriptor extraction, designed a self-training method, supervised learning through a keypoint detector, and realized an end-to-end neural network model for feature point matching. Law and Deng [38] proposed a detection framework based on frame corners, which linked the regression target frame with the feature itself for the first time, pointing out a new direction for target recognition based on feature points and target-based matching tasks. Simo-Serra et al. [39] proposed a discriminant learning method for feature point descriptors. This method uses the Siamese network, takes the nonlinear mapping output by the CNN as the descriptor, and uses the Euclidean distance to calculate the similarity. This method is suitable for processing datasets that contain different types of data and for different applications, including rotation scaling, nonrigid transformation, and illumination change. The Siamese network, proposed by Chopra et al. [40] in 2005, is characterized by two or more subnetworks that simultaneously receive two images as input and share the weights of the two neural networks. More recently, the Siamese network has been widely used for semantic classification and object tracking because of its excellent structural characteristics and relatively simple principle, which makes it suitable for addressing the "similarity problem." If the branch networks on the left and right taken by the two inputs are different, or the network does not share network weights, it is referred to as a pseudo-Siamese network, which is suitable for processing image pairs with certain differences in input. Zbontar and Lecun [41] introduced a Siamese network to calculate the matching cost. The network was trained to predict the similarity between image patches and applied to disparity estimation, similar to converting the calculation of the matching cost to a multilabel classification problem. Han et al. [42] used a patch-level-based Siamese network for feature extraction and matching by using the similarity measure. Balntas et al. [43, 44] proposed the triplet network structure, an extension of the Siamese network, and simultaneously considered the relationship between three samples during training. Zagoruyko and Komodakis [45] improved the Siamese network and extracted the features of two images for comparison, but only the similarity of the two images could be obtained. Contrastingly, the position of the target in the reference image could not be determined. Positioning required a traversal operation, and the real-time performance of their method was poor. Wu et al. [46] used a pseudo-Siamese network to estimate the 6D pose motion of textureless objects, thereby extending the application of the Siamese network. Although the Siamese network has been greatly improved in the past two years, it is basically designed for target tracking [47, 48]. At present, commonly used deep learning image matching methods include D2-Net [49], R2D2 [50], SuperGlue [51], Key.Net [52], and AffNet+HardNet [53]. Existing algorithms are unsuitable for use in any of these methods to meet the requirements of real-time performance and robustness under all conditions. Owing to the scarcity of training samples and real-time requirements, certain algorithms such as those that rely on CNNs and deep learning have found limited application in the field of engineering. In particular, in the case of a drone that uses a camera as the carrier for image matching, the target to be identified is often determined before it is detected and recognized. However, the target position is locked in the reference image. If the target position is only determined by a commonly used image-matching method, unsuccessful matching resulting from the vibration of the drone, changes in the attitude, an excessive viewing distance, and illumination effects could occur, exacerbated by low target resolution and notable defects such as low contrast, distortion, zoom, and lack of texture. In feature-based image-matching methods, matching two images using feature point descriptors may generate incorrect matching points, thus affecting the visual localization effect. Therefore, a method is needed to screen the image-matching results to judge the quality of matching point pairs, eliminate incorrect matching point pairs more accurately, and improve the reliability and accuracy of visual positioning. Contrastingly, the use of a deep learning method for target recognition would result in the target feature information provided by the reference image not being fully utilized, resulting in the feature information of the target being detected. This information must be stored by the target recognition algorithm, which would substantially expand the network structure and increase the computational load. It cannot meet the real-time requirements of embedded systems. A practical image-matching algorithm that is insensitive to influencing factors such as deformation, rotation, and imaging angle changes in the scene is therefore necessary. Particularly under the premise of a small number of samples, using deep learning to perform feature matching operations has been a great test for the generalization ability of the network.

To solve the above problems, we focus on an integrated localization algorithm of target recognition and matching. We combine the single-shot multibox detector (SSD) and deep feature matching (DFM) algorithms to propose an integrated network structure for target recognition and image matching based on the idea of a pseudo-Siamese network. The SSD target recognition algorithm is introduced to extract the target feature information in real time, and reference images are used to select the adaptation area. Based on the trained network structure, the DFM algorithm is used to complete deep feature matching, and the grid-based motion statistics (GMS) algorithm is used to eliminate incorrectly matching pairs. The experimental results show that the algorithm proposed in this paper has stronger robustness and higher matching accuracy than the traditional matching algorithm. The proposed algorithm can effectively improve the generalization ability of the network while ensuring real-time performance and has practical engineering value.

In summary, the main contributions of this paper are as follows:

(1) Based on the idea of a pseudo-Siamese network, an integrated target recognition and matching localization algorithm, which combines the SSD and DFM algorithms, is proposed. The SSD algorithm is used to select the image-matching adaptation area. By sharing the weight of the neural network, the acquisition ability of target features is enhanced, and the integration of target recognition, image matching, and positioning is realized, thereby improving the matching performance and positioning accuracy of the algorithm

(2) A target matching strategy that combines the DFM and GMS algorithms is adopted to reduce the matching error and incorrect matching point pairs in the DFM algorithm, optimize the matching result, and improve the correct matching rate of the target

(3) The University-Release dataset is used to compare and analyze the performance of the proposed algorithm, and an actual flight test is conducted using an S1000 six-rotor UAV. The results of the analysis and flight test show that the performance improved relative to that of the existing matching algorithm and that the performance met the requirements of visual positioning for UAVs. Our method has certain theoretical and practical reference value

The remainder of this paper is structured as follows. Section 3 describes the problem and preliminary work. Section 4 explains the research methods presented in this paper.

Section 5 presents the experimental results and analysis of the proposed algorithm. Section 6 concludes the paper.

3. Problem Description and Preliminaries

The aerial and reference images collected by UAVs while performing image matching and navigation tasks are generally characterized by their high-definition nature, data intensiveness, and high degree of currentness. Processing reference and aerial images is time-consuming and memory intensive. In practical tasks, we usually focus on the areas in which the target is located and refrain from matching the entire image. Therefore, the development of an integrated localization algorithm for both UAV target recognition and matching is challenging. First, information about the target scene is obtained using the airborne camera. Second, the target recognition algorithm is used to perceive the target area, select the adaptation area, extract the target features, and identify the target. In the adaptation area, the target matching strategy, which combines the DFM and GMS algorithms, is used to complete the matching and positioning task of the UAV. The navigation and positioning algorithm of the UAV is the key component of the flight mission system of the UAV. Our work mainly focuses on the UAV navigation task of image matching. The proposed localization algorithm for UAVs is based on the target recognition algorithm, guaranteed by the GMS algorithm, and integrates the tasks of target recognition and image matching.

With the continuous development of deep learning neural networks, the size of the model of target recognition algorithms and the complexity of these algorithms are increasing, which demands higher performance in terms of the operating speed and integration of hardware processors. The current mainstream ARM architecture processors cannot meet the real-time requirements when performing a large number of operations on unstructured data, which greatly limits the application of deep learning algorithms in the field of engineering. Deep neural network models with high recognition accuracy that have been widely used in recent years mainly include the following: (1) the region-based convolutional neural network (R-CNN) series [54–57]; (2) the you only look once (YOLO) series [58–61]; and (3) the SSD series [62–65]. The SSD algorithm integrates the anchor mechanism of Faster R-CNN and the regression principle of YOLO, which improves both the speed and accuracy. SSD uses a multiscale convolution feature map to predict the target area, outputs a series of discrete, multiscale default of the outer frame coordinates, and uses a small convolution kernel to predict the frame coordinates of candidate boxes and the confidence of each category. Therefore, SSD has both the speed of YOLO and the high accuracy of Faster R-CNN. SSD adopts the VGG-16 model as the backbone network [66], which has been improved and modified to some extent. The specific network structure is shown in Figure 1, where "conv" is a convolution operation. The SSD model accepts images of a fixed size as input, integrates feature maps of different levels, calculates the category and confidence of the predefault bounding box, and finally obtains the target detection result using non-maximum suppression. To achieve effective target detection, the loss function of the network model adopts the weighted sum of the localization loss (Loc) and the category confidence loss (Conf) [67].

$$L(x, c, l, g) = \frac{1}{N}\left(L_{\mathrm{conf}}(x, c) + \alpha L_{\mathrm{loc}}(x, l, g)\right), \tag{1}$$

where N is the number of positive samples in the a priori box; x is the indicating function, which indicates whether the default bounding box matches the real bounding box; c is the predicted value of category confidence; α is the weight coefficient; l is the prediction boundary box; g is the real bounding box. The positioning loss function L_{loc} uses a smooth L_1 function to calculate the loss between l and g. The confidence loss function L_{conf} is calculated by softmax. The specific definitions of the localization loss L_{loc} and the confidence loss L_{conf} are

$$\begin{aligned} L_{\mathrm{loc}}(x, l, g) &= \sum_{i\in\mathrm{Pos}}^{N} \sum_{m\in\mathrm{loc}_i} x_{\mathrm{ij}}^{k} \mathrm{smooth}_{L1}\left(l_i^m - \hat{g}_i^m\right), \\ L_{\mathrm{conf}}(x, c) &= -\sum_{i\in\mathrm{Pos}}^{N} x_{\mathrm{ij}}^{p} \log\left(\hat{c}_i^p\right) - \sum_{i\in\mathrm{Neg}} x_{\mathrm{ij}}^{p} \log\left(\hat{c}_i^0\right), \hat{c}_i^p \\ &= \frac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)}, \end{aligned} \tag{2}$$

where $\hat{g}_j^{\mathrm{cx}} = (g_j^{\mathrm{cx}} - d_j^{\mathrm{cx}})/d_j^w$, $\hat{g}_j^{\mathrm{cy}} = (g_j^{\mathrm{cy}} - d_j^{\mathrm{cy}})/d_j^h$, $\hat{g}_j^w = \log($ $g_j^w/d_j^w)$, and $\hat{g}_j^h = \log(g_j^h/d_j^h)$; d_j^{cx}, d_j^{cy}, d_j^{cw}, and d_j^{ch} contain the location information of the target, m represents the number of feature maps, and l_i^m represents the i prediction box in the m category. $\hat{g}_i^m$ represents the j prediction box in the m category. x_{ij}^k indicates whether the i predicted box matches the j real box concerning the k category. If the box matches, the value is 1; otherwise, it is 0. Pos, loc, and Neg represent a set of positive samples, negative samples, and a set of bounding box coordinate positions, respectively.

4. The Proposed Approach

The proposed algorithm is designed to extract target features based on the SSD algorithm, select the adaptation area for target recognition, and use the depth feature matching method to construct an integrated target recognition and matching location algorithm based on a pseudo-Siamese network. Finally, the GMS algorithm is used to eliminate mismatching point pairs to achieve the integrated network function of target recognition and navigation positioning and achieve accurate positioning for image-matching navigation. Section 4.1 discusses the DFM method in detail. Section 4.2 expounds on the error matching elimination strategy based on GMS. Section 4.3 proposes an integrated network structure for target recognition and matching based on a pseudo-Siamese network and explains the principle and process of the algorithm in detail.

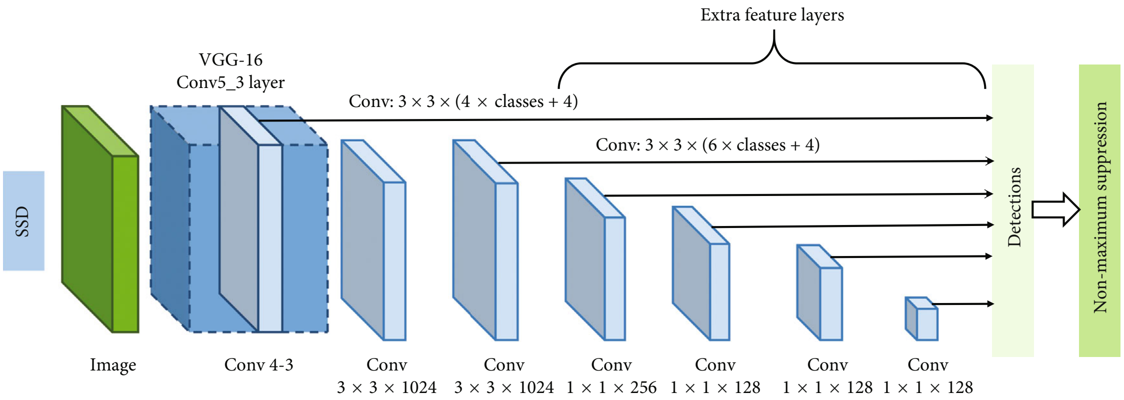

Figure 1: SSD network structure chart.

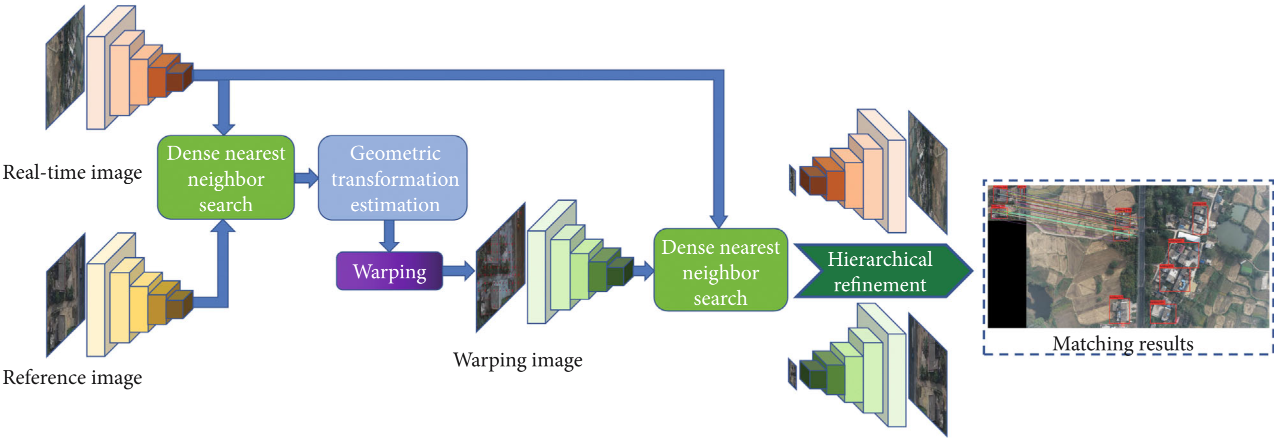
Real-time image
Reference image
Dense nearest neighbor search
Geometric transformation estimation
Warping
Warping image
Dense nearest neighbor search
Hierarchical refinement
Matching results

Figure 2: Process chart of the deep feature matching method.

4.1. Deep Feature Matching Method. DFM is a method that uses the features extracted by deep neural networks to find the point of correspondence between two images. The method does not require external detection and feature description but directly calculates the correspondence between two images [68, 69]. The main idea is to use the VGG network pretrained by the SSD target recognition algorithm to extract features without any special training for feature matching. When using the VGG-16 network for feature extraction, we use the conv5_3, conv5_2, relu4_2, conv3_2, conv2_2, and conv1_2 network layers for feature extraction and use the deepest translation, scale, and brightness invariances of the neural network. Features corresponding to both images are found at the semantic level [70]. The process of the depth feature matching method is shown in Figure 2 [71].

For the given images, A and B, we first use the pretrained VGG-16 network to perform deep feature extraction on these two images to obtain feature blocks F^A and F^B, after which we use the dense nearest neighbor search algorithm to search F^B to find each element of F^A for the best matching position. The dense nearest neighbor search (DNNS) algorithm uses mutual nearest neighbor search and ratio testing in dense feature maps for search matching [72]. Setting the distance L2 as the nearest neighbor matching distance, for point P^A in feature map F^A, point P^A matches P^B if the ratio of distance L2 to P^B of the best match and $P^B - 1$ of the next best match is below a given threshold τ. However, the pair is accepted if they match each other. If P^B also matches P^A, then P^A and P^B are returned as a matched pair. For the matching pair set, the hierarchical refinement method (HRM) is used to implement a coarse-to-fine matching strategy, using the semantic characteristics of the deep network and the detailed characteristics of the shallow network to map the deepest features to the shallowest to achieve accurate matching purpose [73, 74]. $P_n^{A,B}$ represents the matching pair at layer n, while F_{n-1}^A and F_{n-1}^B represent the feature mapping at layer $n-1$. In the VGG-16 network, each point in a matching pair is the parent of 4 points in the previous layer of the network. For each pair of matches, the point sets Ω^A and Ω^B are constructed to represent the receptive fields of P^A and P^B in the $n-1$ layer. We feed the patches of feature maps F^A and F^B to a DNNS algorithm and receive the matched pairs in $n-1$ layers to optimize the matched pairs. We iteratively apply DNNS sequentially using 2×2 feature patches to refine the matching pairs hierarchically by moving to a finer resolution at each step until the first layer. As the features in the shallow layers of the neural network are not as robust to geometric transformations as the deep layers, pairs that have been correctly

```
Input:imageA and imageB
Output:Matching results of two images
// Get the feature map at the $n$ layer
1: imageA->$F_n^A$ and imageB->$F_n^B$
// Get feature map mapping points
2: $F^A$-> $P^A$ and $F^B$->$P^B$
// Get matching pairs at layer $n$
3:   if (L2($P_n^A, P_n^B$)/L2($P_n^A, P_n^B - 1$) < $\tau$)
         $P_n^{A,B} = (P_n^A, P_n^B)$
     else:
         return null
4: Function HRA($F_{n-1}^A$, $F_{n-1}^B$, $P_n^{A,B}$):
   for $P^A$, $P^B$ in $P_n^{A,B}$ do
     (1) Get the receptive fields at layer
$n-1$for feature points defined at layer $n$:
             $\Omega^A$= receptive($P^A$)
             $\Omega^B$= receptive($P^B$)
     (2) Perform Dense Nearest Neighbor Search:
         $M^{A,B} = \text{DNNS}(F_{n-1}^A(\Omega^A), F_{n-1}^B(\Omega^B))$
     (3) Record the matched pair at layer $n-1$:
         $P_{n-1}^{A,B}.append(transform(M^{A,B}))$
     end
   return $P_{n-1}^{A,B}$
// Find the homography matrix from
     matching point pairs
5: $P_{n-1}^{A,B}$ -> $H_{BA}$
6: To initially align imageA and imageB, according to the homography matrix $H_{BA}$, warp the
     imageB to obtain the imageC.
7: imageA and imageC perform steps (2) (3) (4) again.
8: Map the matching points of imageC to imageB by $H_{BA}$.
9: Output matching point pairs of imageA and imageB.
```

Algorithm 1: Depth feature matching method.

matched are usually generated in the deeper network layers. These matching pairs are eliminated as they move to shallower layers during the hierarchical refinement matching process. Therefore, before performing hierarchical refinement matching, P_5^A and P_5^B provided by DNNS are used to obtain a set of matching points. Using this set of matching points, a homography matrix H_{BA} is obtained, and image B is reversed to obtain image C such that images A and B are initially aligned, whereas images A and C are hierarchically refined and matched to find possible matching pairs. Finally, the matching points of image C are mapped to image B through the homography matrix H_{BA} to complete the matching task. An overview of the entire proposed approach is presented in the form of Algorithm 1.

4.2. Mismatch Elimination Strategy Based on GMS. The GMS algorithm is an image-matching algorithm that processes a large number of matching points to accomplish high quality and highly robust image matching based on grid-based feature points as neighborhood support estimators [75, 76]. The principle of the GMS algorithm is illustrated in Figure 3.

In Figure 3, the image on the left is the matching image I_A, the image on the right is the to-be-matched image I_B, and the two images have M and N feature points, respectively. The set of all matching points of I_A and I_B is set to $X = \{x_1, x_2, \cdots, x_{i,} \cdots, x_N\}$, where $x_i = \{M, N\}$ represents a matched pair of feature points. The GMS algorithm converts motion smoothness constraints into statistics. A field with few correct matching points would contain several matching points, whereas a field with no matching points would contain few matching points. Therefore, counting the number of other matching points in the neighborhood of a matching point enables one to judge whether the matching point is correct. According to this characteristic, a feature point M_i exists in area a that matches N_i in area b. The matching pair x_i matches correctly, but the matching pair x_j matches incorrectly. For region a in Figure 3, let s_i denote the x_i neighborhood support estimator, then [77],

$$s_i = |x_i| - 1, \tag{3}$$

where -1 refers to subtracting the matching pair from the total number. Considering that each feature point is independently matched, s_i could be approximately considered

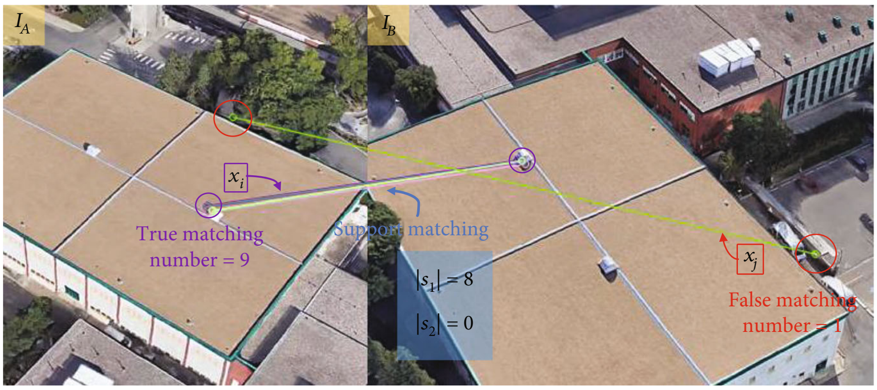

Figure 3: Schematic diagram of GMS principle. In this example, nine pairs of points are correctly matched, whereas one pair is incorrectly matched, and the neighborhood support estimate according to the formula is 8.

to obey a binomial distribution, that is,

$$s_i \sim \begin{cases} B(n, p_t) & \text{if } x_i \text{ is true}, \\ B\left(n, p_f\right) & \text{if } x_i \text{ is false}, \end{cases} \tag{4}$$

Then, the average value m_t and the standard deviation s_t of correctly matching pairs and the average value m_f and standard deviation s_f of incorrectly matching pairs are

$$\begin{cases} m_t = Knp_t, & s_t = \sqrt{Knp_t(1 - p_t)}, \\ m_f = Knp_f, & s_f = \sqrt{Knp_f\left(1 - p_f\right)}. \end{cases} \tag{5}$$

The GMS algorithm typically uses grids of nonoverlapping regions to segment images, and the grid size is set to $G = g \times g$. The s_i value of each matching point pair is calculated in the unit of its grid, thereby reducing the computational complexity of solving the s_i value of each feature point. Figure 4 shows the mesh motion division. The formula for calculating the threshold τ is as follows:

$$\tau = m_f + \alpha s_f, \tag{6}$$

where α represents the adjustment parameter. In practice, the value of m_f is usually small, and the value of α is large. Therefore, the value of τ can be approximately expressed as

$$\tau \approx \alpha s_f \approx \alpha \sqrt{n}. \tag{7}$$

The matching pairs in the grid area in which the neighborhood estimator s_i is greater than τ are retained as the final reliable feature matching pairs.

4.3. PSiamRML Network Architecture. In actual navigation tasks, certain differences in illumination, scale, rotation, translation, deformation, and imaging angle exist between the real image captured by the UAV and the reference image prestored in the navigation system. These changes greatly increase the difficulty and computational complexity of the image-matching task. The image-matching performance is usually improved by focusing on two aspects. The first is to design a more accurate image-matching algorithm. The second involves the selection of a more appropriate adaptation area in the reference image by choosing an adaptation subarea with rich features, good stability, and high significance as the reference image for navigation. During its flight, the UAV could attempt to pass across the adaptation area and bypass the nonadaptation area to realize accurate navigation. With these requirements in mind, based on the SSD target recognition algorithm, our matching strategy consisted of combining the DFM and GMS algorithms to construct an integrated network structure for target recognition and matching based on the pseudo-Siamese network and by sharing network weights. The proposed network structure first selects the adaptation area between the reference and real-time image using the SSD target recognition algorithm. Then, it uses the VGG-16 network structure to extract the target features and the DFM algorithm to match the target in the adaptation area. Finally, it identifies the real-time image in the reference image by selecting its position using the affine transformation box. The network architecture model is shown in Figure 5.

5. Experimental Results and Analysis

We verified the feasibility and superiority of the proposed algorithm by using Opencv3 and MATLAB2016b to conduct related experiments on the University-Release dataset and images recorded by the aerial video recorder of the S1000 six-rotor UAV. The operating system of the ground control station is Ubuntu18.04, and the processor is a laptop with an Intel(R) Core(TM) i7-11800U CPU with a 2.30 GHz processor and 32 GB memory. In this study, nine classical matching algorithms were selected to compare the matching performance, and the positioning effect of the algorithm was analyzed and verified by aerial video recorded by the UAV.

5.1. Comparison of Matching Efficiency. To verify the positioning performance of the PSiamRML algorithm on the same scenes from the University-Release dataset, the

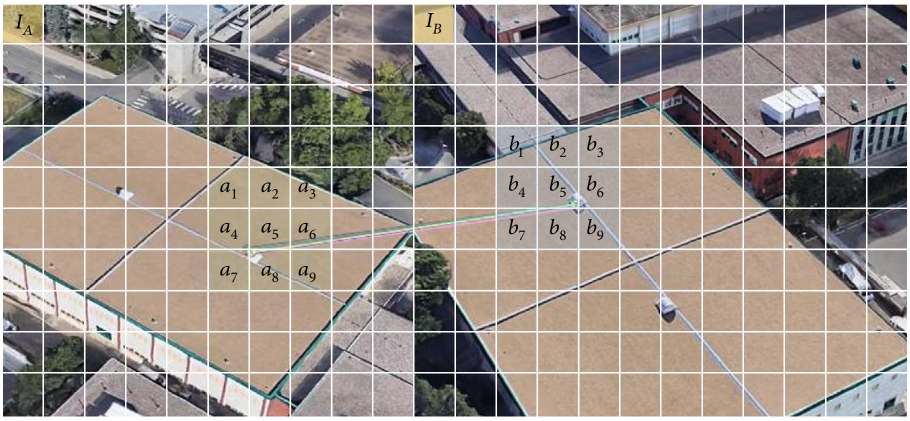

Figure 4: GMS grid motion model.

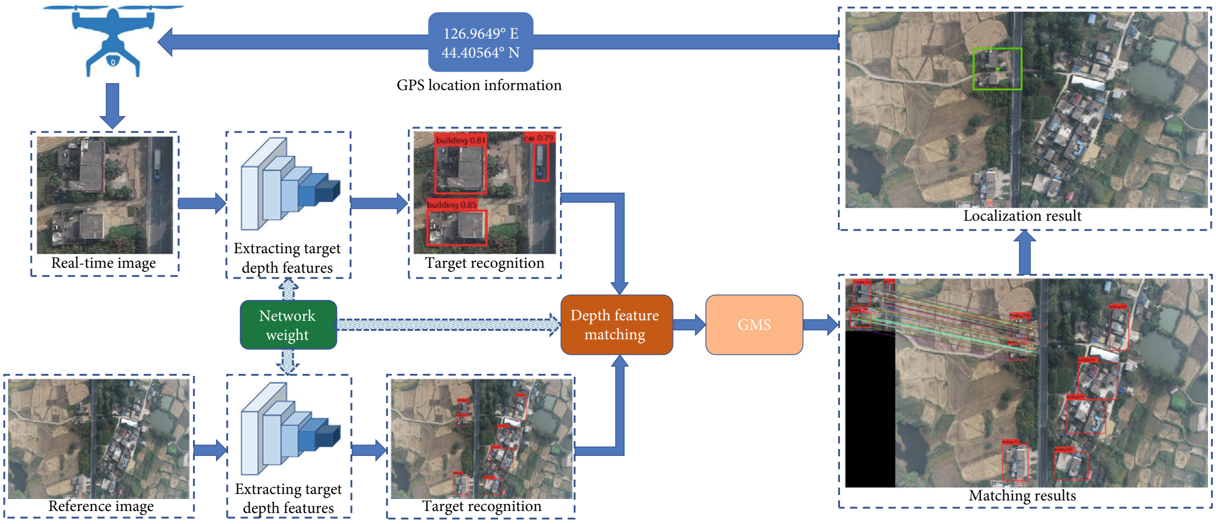

Figure 5: PSiamRML network architecture. PSiamRML algorithm flow: the aerial image captured by the UAV during the image-matching navigation task is matched with the reference map preloaded on the UAV. The algorithm first uses the SSD target recognition algorithm to select the area in which the reference map corresponds with the real-time map, extracts the target features using the VGG-16 network structure, uses the depth feature matching algorithm to match the targets in the corresponding area, and finally selects the position of the real-time map on the reference map using the affine transformation box, and feeds the position navigation information back to the UAV.

significant targets are intercepted from different perspectives as real-time images, and the other perspective images are reference images for positioning verification. In the experiment, variations of four matching scenes with interference effects such as contrast, scale, brightness, blur, and deformation were selected for comparison in the positioning experiment. Among these scenes, scene A is a low-resolution environment at different scales, scene B is an environment with different levels of brightness and blurring, scene C is an environment with different levels of brightness and contrast, and scene D is an environment with different levels of rotation and deformation. The algorithms ORB+GMS, AKAZE+GMS, D2-Net, R2D2, SuperGlue, SuperPoint, AffNet+HardNet, and PSiamRML were used for feature matching. Owing to the different application fields for which image-matching methods are designed, it is difficult to use a unified evaluation index to define the quality of image-matching results [78]. In this study, we analyzed the matching and positioning performance of the algorithm by comparing two metrics: the putative match ratio and positioning error. Assuming that the matching rate refers to the ratio of the number of matching feature points to the total number of features, the calculation formula is as follows:

$$P_{\mathrm{pmr}} = \frac{N_{\mathrm{all}}}{N_F}, \tag{8}$$

where N_{all} is the actual total matching points and N_F is the total number of feature points. The higher the putative match ratio, the greater the number of matching point pairs obtained by the algorithm and the higher the matching performance; however, a few incorrect matching point pairs invariably exist. The positioning error refers to the closeness between the determined UAV position information and its real position based on the relative position relationship of

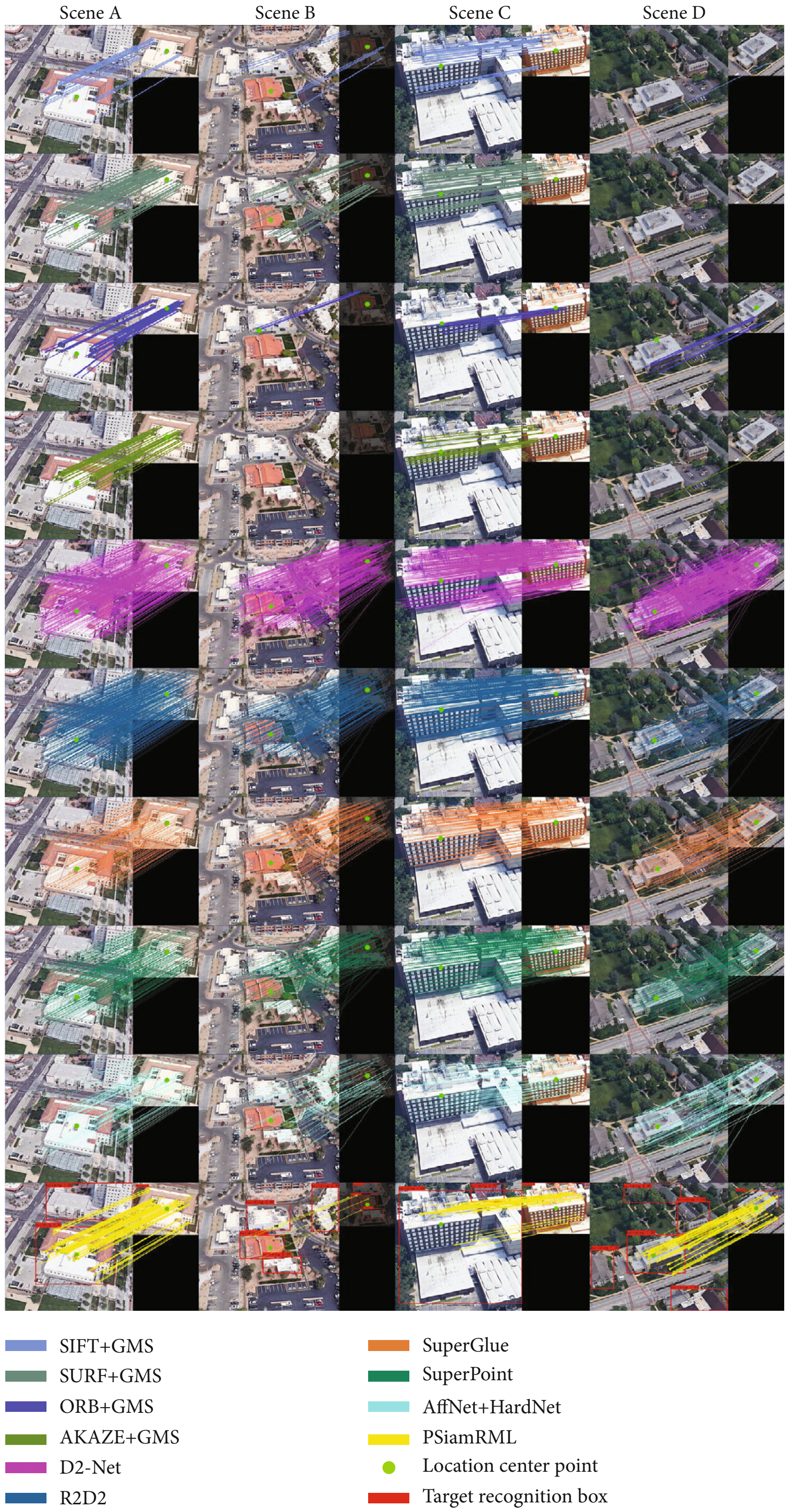

FIGURE 6: Comparison of experimental results of various algorithms for sequences of test images.

the matched feature points. In our work, the L2 distance is used to measure the error in the positioning center between the real-time and reference images. The results of the matching and positioning experiments on the images of the four scenes are shown in Figure 6. The matching results indicate that, under the condition that the GMS algorithm operates effectively, the PSiamRML algorithm has a higher matching success rate without any false matches. In addition, the

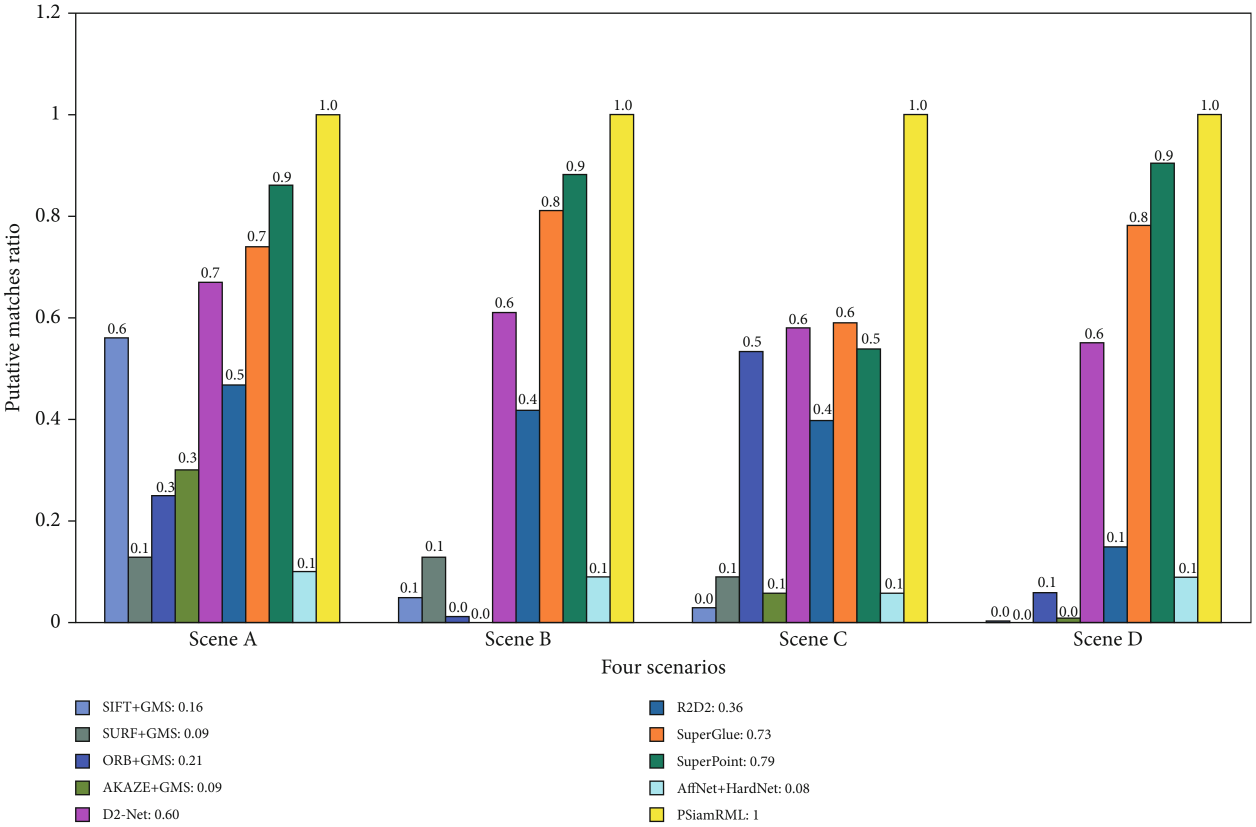

Figure 7: Comparison of putative match ratio of the four typical environments.

algorithm can effectively overcome complex environmental conditions such as scale, blur, deformation, and low resolution. Figures 7 and 8 compare the results of the assumed matching rate and error in the center position in four typical environments to evaluate the matching performance of the PSiamRML algorithm in a complex environment. The legend shows the average value of the assumed matching rate and error in the center position of various algorithms in four environments. Figures 7 and 8 show that, compared with the traditional algorithms SIFT+GMS, SURF+GMS, ORB +GMS, and AKAZE+GMS, PSiamRML is superior in that it identifies the correct matching points, and the positioning error is smaller. Although deep learning matching algorithms such as D2-Net, R2D2, SuperGlue, SuperPoint, and AffNet+HardNet assume a high matching rate and a large number of correct matching points, the positioning error is large because of mismatches. The comparison of the matching efficiency in Figures 7 and 8 shows that the PSiamRML algorithm has high matching accuracy, and it has good stability and robustness against changes in the contrast, scale, brightness, blur, deformation, and so on, combined with a good positioning effect.

5.2. UAV Visual Localization Test. The Pixhawk flight control board was used to independently build the S1000 six-rotor UAV image-matching test system to verify the practicability of the proposed algorithm. The UAV image-matching system includes the UAV, ground control station, remote control, pod, and wireless image transmission equipment. The real-time flight data of the UAV can be transmitted to the ground station in real-time using the digital transmission equipment, and the flight instructions of the ground station can also be transmitted to the UAV. The data transmission equipment uses the 3DRRadio data transmission station V5 module. The frequency is 915 MHz, the transmission power is 1000 MW, and the transmission distance is 5 km. The ground station uses QGroundControl to control drone flight and monitor flight status. When the UAV moves at high speed or performs actions such as pitch, roll, and yaw, it is often subjected to vibration and offset from the external environment and itself, which results in camera shake. The pod is a device that controls the attitude stability of the camera, which prevents the camera from tilting following the UAV, thereby avoiding image instability caused by shaking. The model that was selected for the Pan-Tilt-Zoom (PZD) pod is a TOP-T10XPro. The images collected by the pod are output by an HDMI interface with a resolution of 3840 × 2160 and an output frame rate of 60 FPS. The pod meets the needs of target recognition and matching positioning and ensures real-time image recognition performance. The MK15 model remote control is selected to control the flight of UAV and display real-time visual image information of the pod and UAV inspection area. In image matching, the pod camera captures the real-time image and transmits it to the wireless image transmission equipment, which transmits it to the ground station and remote controller. The designed target recognition and matching integration algorithm on the computer realizes

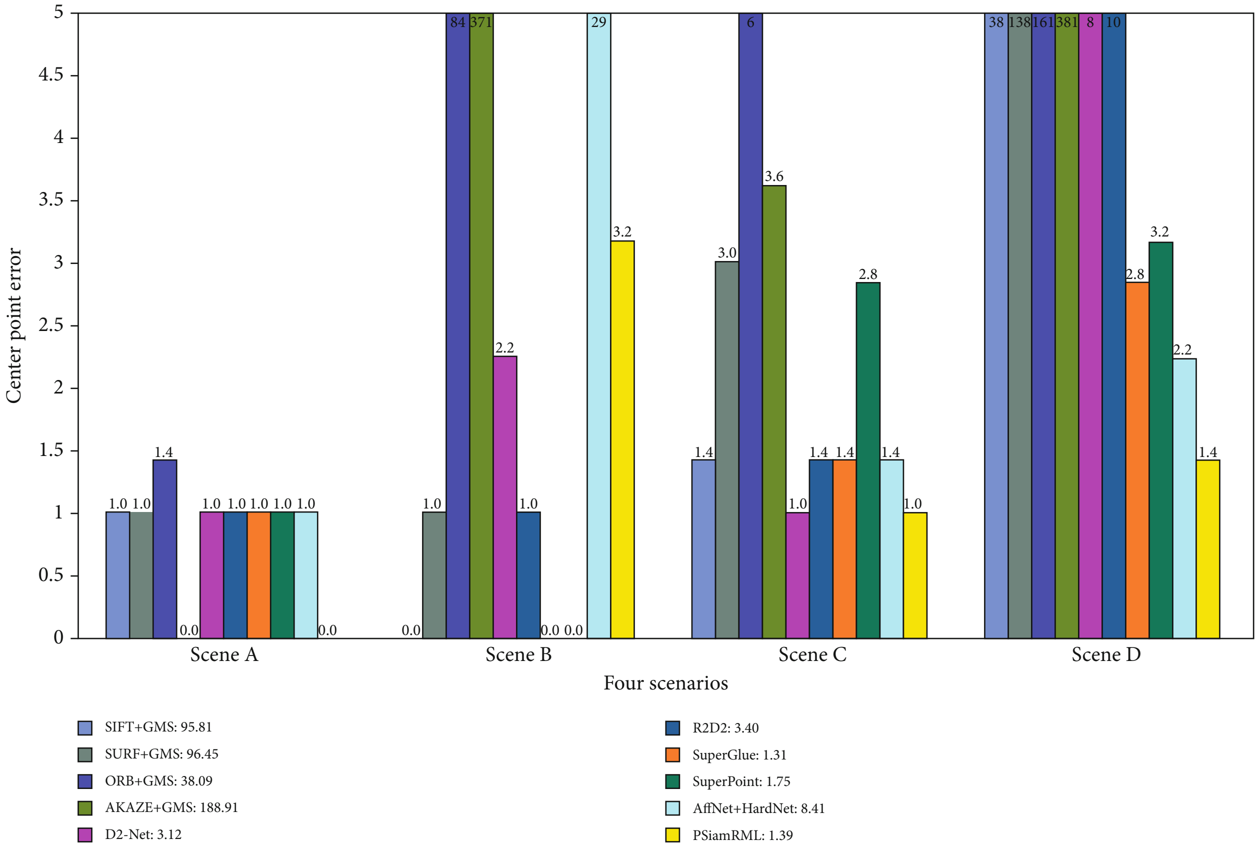

Figure 8: Comparison of center position error of the four typical environments.

target matching and positioning based on the image sequence. Then, the coordinates of the UAV and the target in the world coordinate system are determined by calculating the photographic geometry to adjust the pose of the UAV and realize the visual positioning of the UAV. The configuration of the built UAV image-matching system is shown in Figure 9.

The UAV test scene based on the PSiamRML algorithm is shown in Figure 10 and has various unstructured environmental characteristics. The experimental environment adopts the direct downward flight view of the UAV to more precisely verify the effectiveness of the proposed algorithm. The resolution of the reference image was set to 1200 × 1260, the flying altitude of the UAV was between 300 and 305 m, and the flight distance was 350 m. The video recording was transmitted to the ground control station in real time by the wireless image transmission device and was used as input for the visual positioning algorithm. The aerial image was preprocessed during the experiment, and the resolution of the image collected by the pod was adjusted to 480 × 360 to improve the computing efficiency and save hardware computing resources. The visualization result of the real-time image captured by the UAV is shown in Figure 10. Figure 11 presents the matching result of the UAV matching reference image and target recognition and shows that the matching result of the PSiamRML algorithm completely coincides with the flight trajectory. The PSiamRML algorithm produced a good identification and matching result and could output the corresponding center point of the real-time image in the reference image to realize the navigation and positioning of the UAV.

The test results in Sections 5.1 and 5.2 show that, compared with other matching algorithms, the proposed PSiamRML algorithm delivers greatly improved matching performance and produces good matching and positioning results for images of scenes affected by deformation, blur, and complex backgrounds. Based on the above analysis, the proposed algorithm delivers excellent performance comprehensively with the added advantages of robustness and practicability. Compared with traditional matching algorithms, the accuracy and success rate are also significantly improved. The PSiamRML algorithm has good practical applicability and application value.

5.3. Discussion. Image matching is an important task in the field of computer vision. The advantages of traditional image-matching methods are their simple principles and structures that enable them to run in real-time on a CPU, and their robustness and matching accuracy in general. Image-matching methods based on depth learning use a CNN to extract more robust depth features. Although this has greatly improved the matching accuracy, most depth learning algorithms rely on the powerful computing power of a GPU. In recent years, application of deep learning to image matching has gained attention because a large amount of data is trained using a multilayer CNN to extract the deep

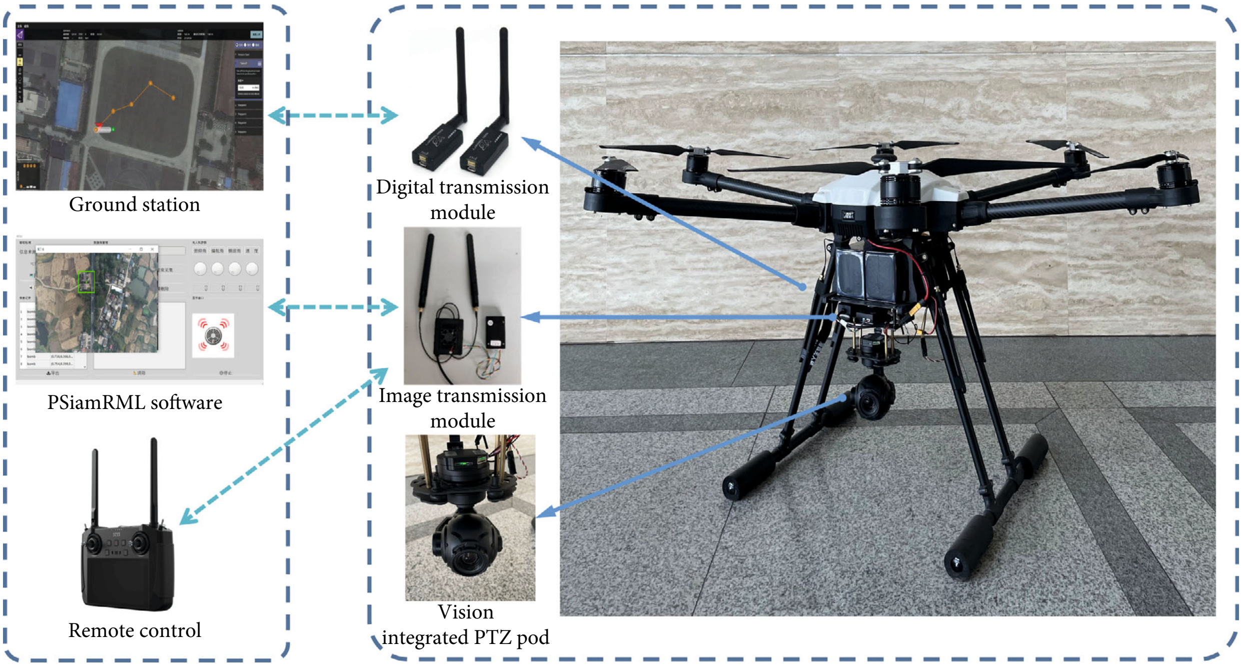
Ground station
PSiamRML software
Remote control
Digital transmission module
Image transmission module
Vision integrated PTZ pod

Figure 9: UAV visual localization test system.

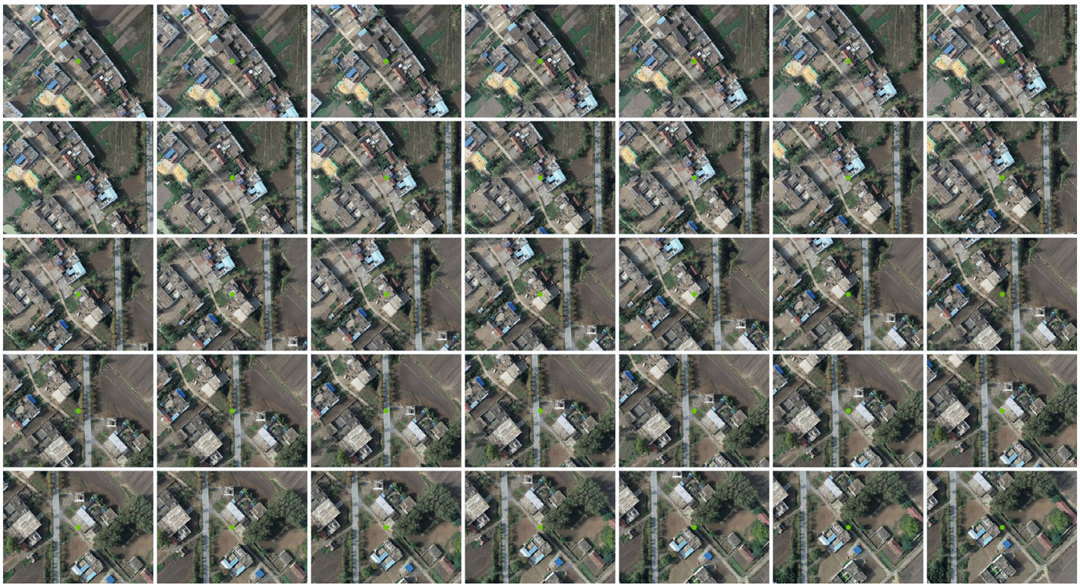

Figure 10: Real-time image captured by UAV ((green circle) geometric center point of real-time image).

features of the target, which further improves the matching effect. However, various problems would need to be solved before the deep learning method would be ready for practical application to image-matching tasks. Currently, the matching region is arbitrary; hence, a network pretrained on an image classification dataset may not be entirely suitable for the image matching and positioning task, which is also a great test for the generalization ability of the deep learning network. The image-matching positioning algorithm needs to have high real-time performance. The deep learning image-matching algorithm improves the robustness of the matching by extracting the deep features of the image using a multilayer CNN, thereby improving the matching and positioning effect. However, an increase in the number of convolutional layers and the complexity of the training network would increase the number of training samples required and the computational intensity. These shortcomings would adversely affect the real-time performance of the algorithm, which would not be conducive to accurately positioning the target. The advantage of the PSiamRML algorithm is that it combines the characteristics of easy implementation of a pseudo-Siamese network with a small number of parameters. It has the ability of a target recognition algorithm to perceive the environment and integrates

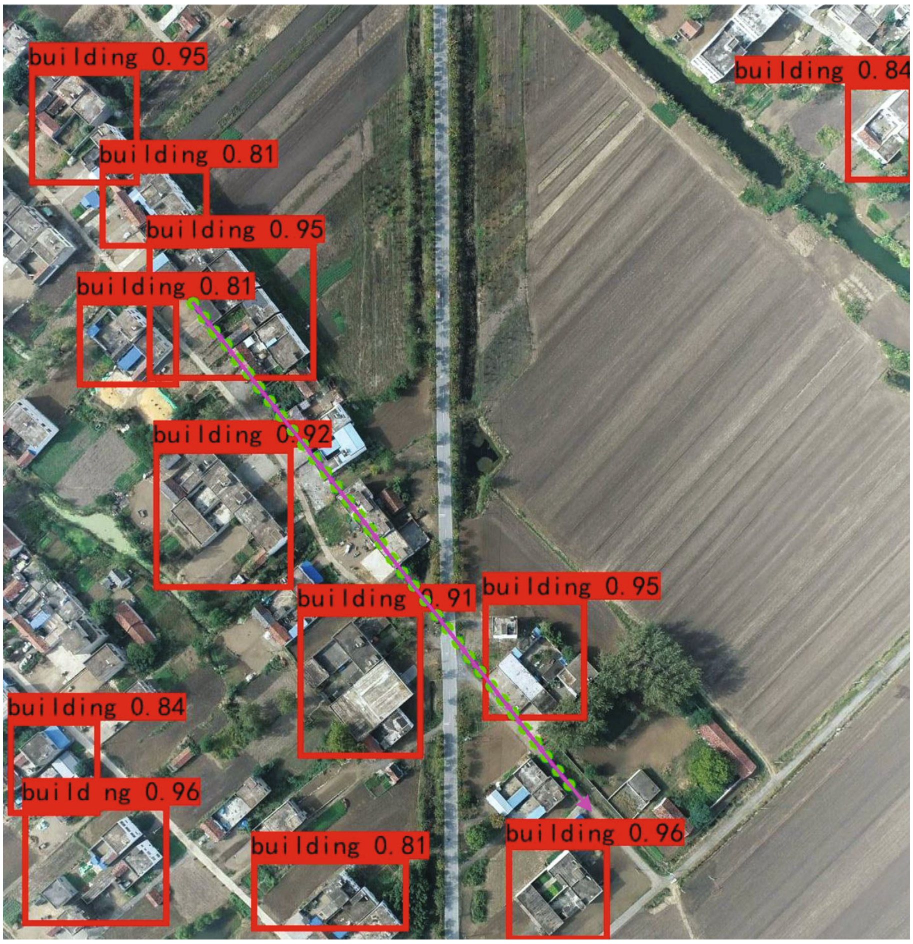

Figure 11: Matching result of the UAV matching reference image and target recognition. ((green circle) center point of real-time image matching; (pink arrow) UAV flight trajectory).

the functions of image matching and positioning of depth features. It effectively integrates target recognition, matching, and positioning and overcomes the deficiencies of a single algorithm.

6. Conclusion

We proposed an integrated localization algorithm for target recognition and matching based on a pseudo-Siamese network. The algorithm adopts the pseudo-Siamese network structure, selects the matching area between the reference and real-time images using the target recognition algorithm, and uses the matching strategy, which combines the depth feature and the GMS algorithm, to complete the target recognition and matching positioning tasks. The experimental results showed that the proposed algorithm has stronger robustness and higher matching accuracy than other matching algorithms. Under the premise of ensuring real-time performance, the generalization ability of the network was effectively improved to realize an integrated design consisting of target recognition and matching positioning algorithms. The computational load is reduced, notably for substantial changes in the illumination, scale, and imaging angle, and the recognition and matching performance are improved. Although the operational efficiency is not as good as that of some deep learning image-matching methods, the PSiamRML algorithm has a simple structure and does not require more prior knowledge of the adaptation area. The superior image-matching effect has practical value in the field of engineering. Future research plans are to improve the operating efficiency of the algorithm while ensuring excellent matching location capability.

Data Availability

The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Disclosure

This paper has been submitted as a preprint, that is, reference [79], and the link is as follows: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4135958.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

X.Y., J.F., and R.L. were responsible for the conceptualization. J.F. and R.L. were responsible for the methodology. J.F. was responsible for the software. X.X. and S.W. were responsible for the investigation. X.X. was responsible for the resources. J.F. and R.L. were responsible for the writing—original draft preparation. X.Y., J.F., and S.W. were responsible for the writing—review and editing. J.F. was responsible for the visualization. J.F. and S.W. were responsible for the supervision. X.Y. was responsible for the project administration. X.Y was responsible for the funding acquisition. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62276274 and in part by the Natural Science Foundation of Shaanxi Provincial Department (2023-JC-YB-528) (corresponding author: Xiaogang Yang.)

References

[1] J. Zhang and W. Zhang, "Cooperative enclosing control with modified guaranteed performance and aperiodic communication for unmanned vehicles: a path-following solution," *IEEE Transactions on Industrial Electronics*, vol. 71, no. 1, pp. 943–953, 2024.

[2] S. Li, X. Shao, W. Zhang, and Q. Zhang, "Distributed multicircular circumnavigation control for UAVs with desired angular spacing," in *Defence Technology*, Elsevier, 2023.

[3] A. E. Hashoosh and M. Alimohammady, "Existence results for nonlinear quasi-hemivariational inequality systems," *Journal of Thi-Qar University*, vol. 11, no. 4, pp. 1–21, 2016.

[4] R. Wang, X. Hou, F. Liu, and Y. Yu, "GPS/INS integrated navigation for quadrotor UAV considering lever arm," in *2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 132–136, Zhanjiang, China, 2020.

[5] H. Hou, Q. Xu, C. Lan et al., "UAV pose estimation in GNSS-denied environment assisted by satellite imagery deep learning features," *IEEE Access*, vol. 9, pp. 6358–6367, 2021.

[6] J. Fan, X. Yang, R. Lu, X. Xie, and W. Li, "Design and implementation of intelligent inspection and alarm flight system for epidemic prevention," *Drones*, vol. 5, no. 3, p. 68, 2021.

[7] M. M. Mostafa, A. M. Moussa, N. El-Sheimy, and A. Sesay, "A smart hybrid vision aided inertial navigation system approach for UAVs in a GNSS denied environment," *Navigation*, vol. 65, no. 4, pp. 533–547, 2018.

[8] X. Shao, S. Li, J. Zhang, F. Zhang, W. Zhang, and Q. Zhang, "GPS-free collaborative elliptical circumnavigation control for multiple non-holonomic vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3750–3761, 2023.

[9] J. Hai, Y. Hao, F. Zou, F. Lin, and S. Han, "A Visual navigation system for UAV under diverse illumination conditions," *Applied Artificial Intelligence*, vol. 35, no. 15, pp. 1529–1549, 2021.

[10] X. Shao, H. Si, and W. Zhang, "Low-frequency learning quantized control for MEMS gyroscopes accounting for full-state constraints," *Engineering Applications of Artificial Intelligence*, vol. 115, article 104724, 2023.

[11] Y. Liu, Y. Gu, J. Li, and X. Zhang, "Robust stereo visual odometry using improved RANSAC-based methods for mobile robot localization," *Sensors*, vol. 17, no. 10, pp. 1–18, 2017.

[12] C. Chi, X. Zhan, S. Wang, and Y. Zhai, "Enabling robust and accurate navigation for UAVs using real-time GNSS precise point positioning and IMU integration," *The Aeronautical Journal*, vol. 125, no. 1283, pp. 87–108, 2021.

[13] W. Youn, H. Ko, H. Choi, I. Choi, J. Baek, and H. Myung, "Collision-free autonomous navigation of a small UAV using low-cost sensors in GPS-denied environments," *International Journal of Control Automation and Systems*, vol. 19, no. 2, pp. 953–968, 2021.

[14] J. C. Trujillo, R. Munguia, S. Urzua, E. Guerra, and A. Grau, "Monocular visual SLAM monocular visual SLAM based on a cooperative UAV-target system," *Sensors*, vol. 20, no. 12, pp. 1–32, 2020.

[15] A. Couturier and M. A. Akhloufi, "A review on absolute visual localization for UAV," *Robotics and Autonomous Systems*, vol. 135, article 103666, 2021.

[16] L. Ding, J. Zhou, L. Meng, and Z. Long, "A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization," *Remote Sensing*, vol. 13, no. 1, pp. 1–20, 2021.

[17] S. Krul, C. Pantos, M. Frangulea, and J. Valente, "Visual SLAM for indoor livestock and farming using a small drone with a monocular camera: a feasibility study," *Drones*, vol. 5, no. 2, p. 41, 2021.

[18] R. Jureviius, V. Marcinkeviius, and J. Eibokas, "Robust GNSS-denied localization for UAV using particle filter and visual odometry," *Machine Vision and Applications*, vol. 30, no. 7-8, pp. 1181–1190, 2019.

[19] S. H. Choi and G. P. Chan, "Image-based Monte-Carlo localization with information allocation logic to mitigate shadow effect," *IEEE Access*, vol. 8, pp. 213447–213459, 2020.

[20] Z. Liu, J. An, and J. Yu, "A simple and robust feature point matching algorithm based on restricted spatial order constraints for aerial image registration," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 50, no. 2, pp. 514–527, 2012.

[21] J. Ma and J. Zhao, "Robust topological navigation via convolutional neural network feature and sharpness measure," *IEEE Access*, vol. 5, pp. 20707–20715, 2017.

[22] M. H. Mughal, M. J. Khokhar, and M. Shahzad, "Assisting UAV assisting UAV localization via deep contextual image matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 2445–2457, 2021.

[23] S. Bas and A. O. Ok, "A new productive framework for point-based matching of airplane oblique and UAV based images," *The Photogrammetric Record*, vol. 36, no. 175, pp. 252–284, 2021.

[24] J. Jin and M. Hao, "Registration of UAV images using improved structural shape similarity based on mathematical morphology and phase congruency," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1503–1514, 2020.

[25] Q. Kai, H. Shi-sheng, and Z. Xi-yang, "On-orbit dynamic scene real-time matching method and experiment of optical satellite," *Chinese Optics*, vol. 12, no. 3, pp. 575–586, 2019.

[26] C. Wei, H. Xia, and Y. Qiao, "Fast unmanned aerial vehicle image matching combining geometric information and feature

similarity," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 12, pp. 1731–1735, 2021.

[27] B. Zhao, H. Wang, L. Tang, and Y. Han, "Towards long-term UAV object tracking via effective feature matching," *Electronics Letters*, vol. 56, no. 20, pp. 1056–1059, 2020.

[28] L. Xing and W. Dai, "A local feature extraction method for UAV-based image registration based on virtual line descriptors," *Signal, Image and Video Processing*, vol. 15, no. 4, pp. 705–713, 2021.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[31] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *2011 International Conference on Computer Vision*, pp. 2564–2571, Barcelona Spain, 2011.

[32] D. Y. Jiang and J. Kim, "Artwork painting identification method for panorama based on adaptive rectilinear projection and optimized ASIFT," *Multimedia Tools and Applications*, vol. 78, no. 22, pp. 31893–31924, 2019.

[33] M. L. Cheng and M. Matsuoka, "An enhanced image matching strategy using binary-stream feature descriptors," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 17, pp. 1253–1257, 2020.

[34] Y. Zhang, J. Song, Y. Ding, Y. Yuan, and H. L. Wei, "FSD-BRIEF: a distorted BRIEF descriptor for fisheye image based on spherical perspective model," *Sensors*, vol. 21, no. 5, pp. 1–26, 2021.

[35] T. Wang, Z. Wang, Y. Cao, Y. Wang, and S. Hu, "A multi-BRIEF-descriptor stereo matching algorithm for binocular visual sensing of fillet welds with indistinct features," *Journal of Manufacturing Processes*, vol. 66, pp. 636–650, 2021.

[36] Y. Dong, D. Fan, Q. Ma, and S. Ji, "Superpixel-based local features for image matching," *IEEE Access*, vol. 9, pp. 15467–15484, 2021.

[37] D. Daniel, M. Tomasz, and R. Adrew, "Super point: self-supervised interest point detection and description," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 224–236, Salt Lake City, 2018.

[38] H. Law and J. Deng, "CornerNet: detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 642–656, 2020.

[39] E. Simo-Serra, E. Trulls, L. Ferraz, L. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of 2015 IEEE International Conference on Computer Vision*, pp. 118–126, Santiago, Chile, 2015.

[40] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 539–546, San Diego, CA, USA, 2005.

[41] J. Zbontar and Y. Lecun, "Stereo matching by training a convolution neural network to compare image patches," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2287–2318, 2016.

[42] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: unifying feature and metric learning for patch-based matching," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3279–3286, Boston, MA, USA, 2015.

[43] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "PN-Net: conjoined triple deep network for learning local image descriptors," 2016, http://arxiv.org/abs/1601.05030.

[44] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks," *British Machine Vision Conference.*, vol. 1, no. 2, 2016.

[45] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4353–4361, Boston, 2015.

[46] C. Wu, L. Chen, Z. He, and J. Jiang, "Pseudo-Siamese graph matching network for textureless objects' 6D pose estimation," *IEEE Transactions on Industrial Electronics*, vol. 6, no. 9, pp. 1–10, 2022.

[47] R. Lu, Y. Xiaogang, W. Li, J. Fan, D. Li, and X. Jing, "Robust infrared small target detection via multidirectional derivative-based weighted contrast measure," *IEEE Geoscience and Remote Sensing Letters*, vol. 1, no. 1, pp. 1–5, 2022.

[48] R. Lu, X. Yang, X. Jing et al., "Infrared small target detection based on local hypergraph dissimilarity measure," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, 2022.

[49] M. Dusmanu, I. Rocco, T. Pajdla et al., "D2- net: a trainable cnn for joint description and detection of local features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8092–8101, California, 2019.

[50] J. Revaud, L. P. Weinzaepfe, C. D. Souza, and M. Humenberger, "R2D2: reliable and repeatable detectors and descriptors," 2019, http://arxiv.org/abs/1906.06195.

[51] P. E. Sarlin, D. Detone, T. Malisiewicz, and A. Rabinovich, "Super glue: learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Virtual, 2020.

[52] A. B. Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.Net: keypoint detection by handcrafted and learned CNN filters," in *Proceedings of the IEEE/CVF international conference on computer vision*, Korea, 2019.

[53] D. Mishkin, F. Radenovic, and J. Matas, "Repeatability is not enough: learning affine regions via discriminability," in *Proceedings of the European conference on computer vision (ECCV)*, Germany, 2018.

[54] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, Columbus, 2014.

[55] X. Ding, Q. Li, Y. Cheng, J. Wang, W. Bian, and B. Jie, "Local keypoint-based faster R-CNN," *Applied Intelligence*, vol. 50, no. 10, pp. 3007–3022, 2020.

[56] J. Qin, Y. Zhang, H. Zhou, F. Yu, B. Sun, and Q. Wang, "Protein crystal instance segmentation based on mask R-CNN," *Crystals*, vol. 11, no. 2, pp. 157-158, 2021.

[57] M. H. Wu, H. H. Yue, J. Wang et al., "Object detection based on RGC mask R-CNN," *IET Image Processing*, vol. 14, no. 8, pp. 1502–1508, 2020.

[58] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pp. 429–442, Las Vegas, 2016.

[59] Y. Xue, Z. Ju, Y. Li, and W. Zhang, "MAF-YOLO: multi-modal attention fusion based YOLO for pedestrian detection," *Infrared Physics & Technology*, vol. 118, article 103906, 2021.

[60] A. Fj, B. Jas, C. Anm et al., "Detection of mold on the food surface using YOLOv5," *Current Research in Food Science*, vol. 4, p. 754, 2021.

[61] X. Li, "A real-time detection algorithm for kiwifruit defects based on YOLOv5," *Electronics*, vol. 10, no. 14, pp. 1711–1713, 2021.

[62] S. Zhai, D. Shang, S. Wang, and S. Dong, "DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion," *IEEE Access*, vol. 8, pp. 24344–24357, 2020.

[63] D. Jia, J. Zhou, and C. Zhang, "Detection of cervical cells based on improved SSD network," *Multimedia Tools and Applications*, vol. 2, pp. 1–17, 2022.

[64] H. Pan, Y. Li, and D. Zhao, "Recognizing human behaviors from surveillance videos using the SSD algorithm," *The Journal of Supercomputing*, vol. 77, no. 7, pp. 6852–6870, 2021.

[65] X. Zhang, H. Xie, Y. Zhao, W. Qian, and X. Xu, "A fast SSD model based on parameter reduction and dilated convolution," *Journal of Real-Time Image Processing*, vol. 18, no. 6, pp. 2211–2224, 2021.

[66] P. Sun, Y. Zhao, and S. Zhu, "An approach to improve SSD through mask prediction of multi-scale feature maps," *Pattern Analysis and Applications*, vol. 24, no. 3, pp. 1357–1366, 2021.

[67] J. Yang, W. Y. He, T. L. Zhang, C. L. Zhang, L. Zeng, and B. F. Nan, "Research on subway pedestrian detection algorithms based on SSD model," *IET Intelligent Transport Systems*, vol. 14, no. 11, pp. 1491–1496, 2020.

[68] I. Rocco, M. Cimpoi, R. Arandjelovic, A. Torii, T. Pajdla, and J. Sivic, "Ncnet: neighbourhood consensus networks for estimating image correspondences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 1020–1034, 2022.

[69] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6148–6157, Hawaii, 2017.

[70] X. Zan, X. Zhang, Z. Xing et al., "Automatic detection of maize tassels from UAV images by combining random forest classifier and VGG16," *Remote Sensing*, vol. 12, no. 18, p. 3049, 2020.

[71] U. Efe, K. G. Ince, and A. Alatan, "DFM: a performance baseline for deep feature matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Virtual, 2021.

[72] T. Patten, K. Park, and M. Vincze, "DGCM-net: dense geometrical correspondence matching network for incremental experience-based robotic grasping," *Frontiers in Robotics and AI*, vol. 7, pp. 1–39, 2020.

[73] J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "DeepMatching: hierarchical deformable dense matching," *International Journal of Computer Vision*, vol. 120, no. 3, pp. 300–323, 2016.

[74] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: CNNs for optical flflow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, Salt Lake City, 2018.

[75] C. Li, B. Guo, X. Guo, Y. Zhi, J. Tseng, and I. Kotenko, "Real-time UAV imagery stitching based on grid-based motion statistics," *Journal of Physics: Conference Series*, vol. 1069, article 012163, 2018.

[76] Q. Tang, J. Yang, W. Jia, X. He, Q. Zhang, and H. Liu, "A GMS-guided approach for 2D feature correspondence selection," *IEEE Access*, vol. 8, pp. 36919–36929, 2020.

[77] J. W. Bian, W. Y. Lin, Y. Liu et al., "GMS: grid-based motion statistics for fast, ultra-robust feature correspondence," *International Journal of Computer Vision*, vol. 128, no. 6, pp. 1580–1593, 2020.

[78] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615–1630, 2015.

[79] F. Jiwei, X. Yang, R. Lu, X. Xie, and S. Wang, "Psiamrml: target recognition and matching integrated localization algorithm based on pseudo-Siamese network," 2022 https://ssrn.com/abstract=4135958.