Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.039582

ARTICLE

Multi-Modal Scene Matching Location Algorithm Based on M2Det

Jiwei Fan, Xiaogang Yang*, Ruitao Lu, Qingge Li and Siyu Wang

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

*Corresponding Author: Xiaogang Yang. Email: doctoryxg@163.com

Received: 06 February 2023 Accepted: 14 August 2023 Published: 31 October 2023

ABSTRACT

In recent years, many visual positioning algorithms have been proposed based on computer vision and they have achieved good results. However, these algorithms have a single function, cannot perceive the environment, and have poor versatility, and there is a certain mismatch phenomenon, which affects the positioning accuracy. Therefore, this paper proposes a location algorithm that combines a target recognition algorithm with a depth feature matching algorithm to solve the problem of unmanned aerial vehicle (UAV) environment perception and multi-modal image-matching fusion location. This algorithm was based on the single-shot object detector based on multi-level feature pyramid network (M2Det) algorithm and replaced the original visual geometry group (VGG) feature extraction network with the ResNet-101 network to improve the feature extraction capability of the network model. By introducing a depth feature matching algorithm, the algorithm shares neural network weights and realizes the design of UAV target recognition and a multi-modal image-matching fusion positioning algorithm. When the reference image and the real-time image were mismatched, the dynamic adaptive proportional constraint and the random sample consensus consistency algorithm (DAPC-RANSAC) were used to optimize the matching results to improve the correct matching efficiency of the target. Using the multi-modal registration data set, the proposed algorithm was compared and analyzed to verify its superiority and feasibility. The results show that the algorithm proposed in this paper can effectively deal with the matching between multi-modal images (visible image–infrared image, infrared image–satellite image, visible image–satellite image), and the contrast, scale, brightness, ambiguity deformation, and other changes had good stability and robustness. Finally, the effectiveness and practicability of the algorithm proposed in this paper were verified in an aerial test scene of an S1000 six-rotor UAV.

KEYWORDS

Visual positioning; multi-modal scene matching; unmanned aerial vehicle

1 Introduction

With the outbreak of many military wars, the human war form has changed from information to an intelligent form. In combat tasks, higher requirements have been put forward for the precision attack capability of weapons and equipment, and the precision positioning and anti-jamming capability of guidance systems. The degree of intelligence of weapons and equipment changes the combat style and methods of future wars. Unmanned warfare has become an important development

trend [1]. Missiles, unmanned aerial vehicles, and other types of aircraft require a navigation system to constantly determine their position to adjust the operating state when performing tasks; therefore, it is very important to further study UAV navigation systems [2]. At present, navigation systems can be divided into autonomous navigation and non-autonomous navigation according to the degree of dependence on external information. Autonomous navigation means that it does not rely on manually set information sources, but only uses its equipment to achieve accurate navigation, such as inertial navigation, visual navigation, and doppler navigation. Non-autonomous navigation refers to technologies that rely on receiving external information for navigation and positioning, such as satellite navigation, radar navigation, and radio navigation [3–5]. Due to the special working environment of aircraft, the main navigation methods used today are global satellite navigation, inertial navigation, and visual navigation. The traditional satellite navigation mode is greatly affected by the environment, and the flight safety of the aircraft faces a huge threat if the satellite signal is interfered with, deceived, etc. Inertial navigation relies on the inertial navigation components carried by the aircraft itself to complete the navigation task. It is not easily interfered with by external information when it is working, but with the extension of the working time, there will be serious pose drift. Therefore, the aircraft's navigation information must be corrected at regular intervals during use [6]. Visual navigation refers to the use of computer vision, image processing, and other technologies to obtain the spatial information and motion information of UAVs. Common technologies can be divided into two categories according to whether prior knowledge is used. One of these categories includes the use of UAV aerial image sequences for matching to obtain the position and attitude transformation relationship of the UAV. Mature technologies include visual odometry (VO) and simultaneous localization and mapping (SLAM) [7,8]. SLAM technology can realize self-positioning while building maps using real-time images in unknown environments. It is widely used in various indoor positioning scenes, but the positioning effect is poor in open outdoor scenes. VO uses UAV aerial images for inter-frame matching to calculate the position and attitude transformation relationship of UAVs. Its principle is similar to that of the inertial navigation system, and the positioning results will also have a large deviation as time goes by. The other category is image-matching navigation, which uses the UAV to select the ground object scene in the predetermined flight area as the reference image database. When the UAV reaches the predetermined area, the airborne camera acquires the current scene in real-time as a real-time image and sends it to the airborne computer for matching and comparison with the reference image in the database. According to the matching position of the real-time image, the current position of the UAV can be determined [9,10]. Image-matching navigation is an absolute positioning technology to realize UAV navigation, which can provide an accurate positioning guarantee for long-endurance UAVs. Image-matching technology was first applied to the terminal guidance of cruise missiles, which has the characteristics of strong autonomy, simple equipment structure, and high positioning accuracy, and has gradually developed into a visual navigation technology. Image-matching technology is to align multiple images acquired by different sensors, different angles, and different time phases in space to determine the relative position relationship between images in the same coordinate system. Its main purpose is to search for the best matching position of the real image in the reference image and to provide basic data for the change in carrier position [11]. In the navigation task, to provide the UAV with the ability to work all day and in all weather conditions, the multi-modal remote sensing image-matching method is generally used for navigation and positioning [12,13]. Multimodal remote sensing image-matching refers to image-matching between different imaging methods (such as satellite images and aerial images) and different sensors (such as visible and infrared images, and visible and satellite images). One or more differences in the imaging characteristics, scale, angle of view, geometry, and other aspects of multimodal image-matching, bring great difficulties to the matching work. At present,

multimodal image-matching methods are mainly divided into gray-based image-matching methods and feature-based image-matching methods [14–17].

In dealing with the actual UAV image-matching and positioning problem, it is necessary to consider the working conditions of the UAV under all weather conditions. Therefore, the high-resolution optical image or satellite image is usually used as the reference image for image-matching. The real-time image mostly uses an infrared image or synthetic aperture radar (SAR) image to meet the all-weather working requirements of the navigation system. In feature-based image-matching methods, feature extraction is a very important component. Traditional feature extraction methods mainly include scale-invariant feature transform (SIFT) [18], oriented fast and rotated brief (ORB) [19], features from accelerated segment test (FAST) [20], histogram of orientated gradient (HOG) [21], affine-SIFT (ASIFT) [22], binary robust invariant scalable keypoints (BRISK), binary fisheye spherical distorted robust independent elemental features (FSD-BRIEF) [23], etc. Because traditional feature extraction methods do not fully utilize data, they can only extract certain aspects of image features. As a result, most of these methods only applying to image-matching tasks in specific scenes. In the consideration of multimodal image-matching, the characteristics of the same feature descriptor in different images are often quite different due to different imaging mechanisms, which makes it difficult to ensure the reliability of the matching results [24]. With the rapid development of depth learning, many feature-matching methods based on convolutional neural networks (CNN) and generic adversarial networks (GAN) have been proposed. Compared with traditional image-matching methods, depth learning methods can learn the shape, color, texture, semantic level, and other features of images through training models, and rotation has a certain invariance. Compared with traditional feature extraction methods, it has stronger description ability and higher generalization. It does not require the manual design of complex features, does not rely on the prior knowledge of designers, and can be effectively generalized to image pairs of other modes [25–29]. In recent years, more and more deep learning feature extraction networks have been proposed, mainly including VGG [30,31], residual network (Res Net) [32,33], dense convolutional network (Dense Net) [34,35], dual path network (DPN) [36,37], neural architecture search network (NASNet) [38], etc.

Recently, some research has been carried out on the image-matching of depth learning. Reference [39] extracted features from image regions through CNN and used a metric network composed of three full connection layers to calculate the matching score of feature pairs, which can effectively reduce the computation of the feature network and improve the accuracy of image-matching. Reference [40] proposed a local feature extraction network with global perception (point pair feature network, PPFNet), which has rotation invariance, can extract local features of 3D point clouds, and can make full use of the sparsity of point clouds, and improve the recall rate, but PPFNet takes up a large amount of memory, so the application effect is poor in some specific tasks. Reference [41] drew on the idea of a transformer and used the self-attention layer and mutual attention layer to obtain the feature descriptors of two images. This method can produce dense matching in areas with less texture. The end-to-end network structure can optimize the whole process of feature matching through information feedback during training; however, the feature descriptors learned by using such methods alone cannot guarantee the matching effect. Reference [42] and others proposed a feature point descriptor discrimination learning matching method, which uses a Siamese network, takes the nonlinear mapping of the CNN output as the descriptor, and uses Euclidean distance to calculate the similarity. This method is applied to different data sets and applications, including rotation scaling, non-rigid transformation, and illumination variation. The Siamese network is characterized by being composed of two or more sub-networks, and the weights of the two neural networks are shared. The Siamese network is suitable for dealing with "similarity problems" because of its excellent structural

characteristics and simple principle. In recent years, it has been widely used in the field of semantic classification and target tracking. If the left and right branch networks adopt different network structures or do not share network weights, they are called pseudo-Siamese networks. The pseudo-Siamese network can extract the common features of multimodal images through different network models, which can effectively alleviate the nonlinear image differences between images. Reference [43] improved the Siamese network by extracting the features of two images for comparison. Nevertheless, only the similarity between the two images can be obtained, not the target's location in the reference image. To achieve positioning, traversal operations are required, but real-time performance is poor. Due to the different imaging mechanisms of multimodal images, it is difficult to directly apply the traditional homologous image-matching algorithm to multimodal image-matching. To reduce the difference between multimodal images, many scholars have transformed images of different modes into images of the unified mode through style transfer learning to eliminate imaging differences between images of different modes. Reference [44] proposed a visible–infrared image-matching method based on generating a countermeasure model centered around the idea of generating countermeasure network transformation and traditional local feature extraction. This method reduces the difficulty of heterogeneous image-matching and provides a new idea for multimodal image-matching, but the matching accuracy of this method is limited by the completeness of image conversion model samples [45]. At present, the commonly used deep learning image-matching methods are D2-Net [46], R2D2 [47], SuperGlue [48], Key.Net [49], PN-Net [50], CFOG [51], RIOS [52], AffNet+HardNet [53], and so on. The existing algorithms cannot meet the requirements of robustness and operation in real time under any conditions. It is not possible to apply some algorithms, such as convolution neural networks and deep learning, to engineering since there are no training samples and they are not available in real-time. In particular, in the UAV image matching based on the camera, when the commonly used image matching methods determine the target position, not only the obvious defects of low target resolution, low contrast, distortion, zoom and so on should be considered. In addition, consideration should also be given to matching failures caused by drone vibration, posture changes, angle of view changes, lighting, and lack of texture. In the feature-based image-matching method, matching two images through feature point descriptors may produce wrong matching points, which will affect the visual positioning effect. Therefore, a method of screening the image-matching results is needed to judge the advantages and disadvantages of matching point pairs, to better eliminate mismatching point pairs, and improve the reliability and accuracy of visual positioning. A practical image-matching algorithm requires insensitivity to the factors existing in the scene, such as imaging characteristics, geometric deformation, scale change, rotation change, and so on. Furthermore, especially under the premise of a small number of samples, determining how to use deep learning for a feature-matching operation is a great test for the generalizability of the network.

This paper proposed a fusion location algorithm for recognizing targets and matching multi-modal scenes to solve the above problems. We combined the M2Det with the depth feature matching algorithm to propose an integrated network structure of target recognition and image matching. The target features of the real-time image and the reference image were extracted by the improved M2Det target recognition algorithm, the depth feature matching was completed based on the trained network structure, and the dynamic adaptive Euclidean distance and random sample consensus (RANSAC) consistency algorithm were used to eliminate the error matching. The experimental results showed that the algorithm proposed in this paper has stronger robustness and higher matching accuracy than the traditional matching algorithm, and can effectively deal with the matching difficulties caused by different imaging modes, resolutions, and scales of multimodal remote sensing images while ensuring it is carried out in real-time. This can effectively improve the generalization ability of the network.

Compared with the target recognition algorithm, the target recognition rate of the algorithm proposed in this paper was higher and has a certain engineering practical value.

The main contributions of this paper are as follows:

• A target recognition matching fusion localization algorithm is proposed, which combines the M2Det algorithm with the depth feature matching algorithm. The feature extraction network of the M2Det algorithm was improved to enhance the feature acquisition capability of the target environment. By sharing the neural network weight, the UAV target recognition and image-matching positioning algorithm were integrated, to improve the matching performance and positioning accuracy of the algorithm, and solve the functional defects of a single algorithm.

• The target matching strategy combines the deep feature based Brute-force (BF) matcher matching algorithm with the dynamic adaptive Euclidean distance RANSAC consistency algorithm. By using this method, we can reduce matching errors and incorrect matching point pairs in the algorithm, optimize matching results, and increase the correct matching rate.

• An algorithm presented in this paper was compared and analyzed in the multi-modal registration data set as well as tested in a real-life six-rotor UAV flight. The analysis and test results showed that the performance of the algorithm proposed in this paper was improved compared with the existing matching algorithm, and it can meet the requirements of UAV visual positioning and has certain theoretical and practical reference values.

The structure of this paper is as follows: Section 2 describes the problems and preparations. Section 3 describes the research methods of this paper. Section 4 gives the experimental results and analysis of the proposed algorithm. Section 5 presents the conclusion.

2 Problem Description and Preliminaries

To enable UAVs to fly autonomously all day, and in all weather conditions, first of all, UAVs must have persistent and stable scene perception and motion perception. Aerial images acquired by UAVs in the process of multi-modal image-matching navigation tasks are generally characterized by a high resolution, a large difference in imaging characteristics, high visibility, etc. It requires a lot of time and computing memory to process reference images and aerial images, and it is difficult to achieve the unification of the UAV environment perception and navigation positioning methods, Therefore, it is a very difficult task to develop a UAV target recognition matching fusion localization algorithm. This task first obtains the scene information of the UAV flight through the airborne camera; then, the target recognition algorithm is used to perceive the flight area, extract the target features and identify the target of interest, and completes the UAV matching and positioning task in the flight area based on the target matching strategy, combining the depth feature matching algorithm and the dynamic adaptive Euclidean distance random consistency algorithm. The navigation and positioning algorithm of UAVs is key for UAVs to be able to perform flight tasks. This paper mainly focuses on the UAV image-matching navigation task, based on the target recognition algorithm, and guaranteed by the dynamic adaptive Euclidean distance random consistency algorithm, which studies the fusion positioning algorithm for UAV target recognition and image-matching.

With the rapid development of deep learning, the model complexity of the target recognition algorithm is increasing, and the memory footprint is also getting larger and larger, which puts forward higher requirements for the integration and operation speed of hardware processors. The current mainstream advanced reduced instruction set computer machine (ARM) architecture and some edge computing processing devices, at this stage, cannot meet the real-time requirements when

processing a large number of unstructured data; consequently, the application of some deep learning neural network algorithms in engineering is limited. In recent years, deep neural networks with high recognition accuracy and wide application mainly include (1) you only look once (YOLO) series [54–56]; (2) the region-based volatile neural network (R-CNN) series [57–59]; and (3) the single-shot multi-box detector (SSD) series [60–62]. The SSD algorithms series combines Faster R-CNN's anchor mechanism with YOLO's regression idea; therefore, SSD algorithms have both the characteristics of Faster R-CNN's high accuracy and the rapidity of YOLO. The principle of the SSD algorithm is to use the VGG16 network to extract features, detect and classify feature maps with different scales, generate multiple candidate boxes, and finally generate detection results through non-maximum suppression steps. However, the backbone network used in this method can only classify targets. Due to the small number of shallow feature convolution layers and insufficient shallow feature extraction, the SSD algorithm has a poor detection effect on small and weak targets. To solve this problem, Zhao et al. proposed the M2Det algorithm based on the multi-level feature pyramid network (MLFPN) structure [63]. This algorithm uses VGG16 as the backbone feature extraction network and integrates MLFPN based on the SSD model. Compared with the SSD model, the combined M2Det backbone feature extraction network and MLFPN extract features from the input image generate dense boundary boxes and category scores according to the learned features, and use non-maximum inhibition to predict, to obtain the final prediction result. The MLFPN network adopted by M2Det inherits the advantages of the feature pyramid network feature extraction network in the SSD algorithm and refines the size information of the target. The MLFPN network module is mainly composed of three parts: feature fusion module (FFM), thinned U-shape module (TUM), and scale-wise feature aggregation module (SFAM). In the MLFPN network, firstly, the features extracted from the backbone network are aggregated into a base feature with richer semantic information through the FFMv1 module. Then, the two largest effective feature layers generated by the TUM module are fused through the FFMv3 module. The results of the FFMv3 fusion and the basic features are fused through the FFMv2 module to obtain multi-level and multi-scale features. Finally, SFAM stacks multi-level features obtained from TUM according to different dimensions, applies an adaptive attention mechanism, forms a multi-level feature pyramid, and generates boundary boxes and category scores with uneven confidence. Finally, the non-maximum suppression (NMS) prediction network is used to remove the boundary boxes with low confidence, to obtain the prediction results closest to the target object. The network structure of the M2Det algorithm is shown in Fig. 1 [64,65].

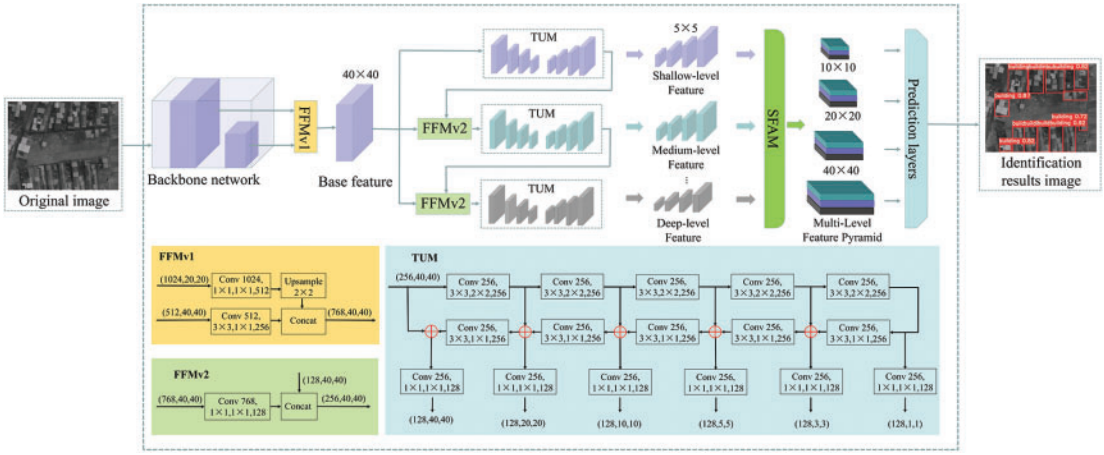


Figure 1: M2Det algorithm network structure

3 The Proposed Approach

The basic idea of the multi-modal scene matching location algorithm based on M2Det (MCML) proposed in this paper is to extract the real-time aerial image features of UAVs based on the improved M2Det network and perceive the real-time image feature information through the target recognition algorithm. A fusion positioning algorithm of target recognition and multi-mode scene matching based on depth features is constructed. Finally, a dynamic adaptive European range random consistency algorithm is used to eliminate mismatched point pairs, realizing the integrated network function design of UAV target recognition and navigation positioning, and achieving the goal of image-matching navigation and accurate positioning. Section 3.1 proposes the network structure of the multi-mode scene matching positioning algorithm based on M2Det, and details the algorithm's principle process. Section 3.2 describes the improved M2Det target recognition algorithm. Section 3.3 discusses the depth feature matching method in detail. Section 3.4 describes the error-matching elimination strategy based on dynamic adaptive Euclidean distance random consistency.

3.1 MCML Algorithm Process

In the vision navigation task of UAV image-matching, the main function of the image-matching algorithm is to realize the navigation and positioning of the UAV. However, the image-matching algorithm cannot obtain the visual situation awareness information of the UAV, and there are changes in illumination, rotation, translation, and affine between the UAV aerial real-time image and the reference image pre-stored by the navigation system, which greatly increases the difficulty of the image-matching task. The main function of the UAV target recognition algorithm is to judge whether there are interesting targets in the image and mark their respective categories and positions in the image, to realize the situation awareness function of the UAV to the flight environment; however, the target recognition algorithm cannot output the navigation position of the UAV in the geography. At present, the UAV image-matching algorithm and target recognition algorithm are studied separately and independently, and an integrated theoretical system has not been formed yet. Therefore, it is also urgent to build a fusion positioning algorithm for UAV image-matching and target recognition, complete the autonomous positioning and target recognition of UAVs in the complex environment without satellites and unknown environments, and give consideration to the real-time nature and accuracy of both. Therefore, based on the improved M2Det target recognition algorithm, this paper introduces the depth feature matching algorithm and the error matching elimination strategy of dynamic adaptive Euclidean distance random consistency. By sharing the neural network weights, this paper constructs the fusion localization algorithm of target recognition and multi-modal scene matching based on the depth feature. The network structure proposed in this paper firstly used the improved M2Det algorithm to train the flight scene, perceives the flight environment situation information by extracting the image features of the UAV aerial real-time image and the reference image, and at the same time, used the depth feature matching algorithm to match the images. Finally, the position of the real-time image in the reference image was selected through the affine transformation box to achieve the accurate positioning of the UAV. The MCML algorithm flow is shown in Fig. 2.

3.2 Improved M2Det Target Recognition Algorithm

The original M2Det target recognition algorithm uses VGG as the backbone feature extraction network. As it consumes more computing resources and takes up more memory space, the general lightweight network improves the detection speed, but the detection accuracy also decreases. Classical deep learning feature extraction models mainly include the VGG series, Inception series, and ResNet series. In terms of model structure, VGG series models have a shallow network depth, Inception series

models have a complex network structure, and ResNet series models have a simple structure. When the number of network layers is deepened, it can solve the problem of gradient disappearance and has strong performance and excellent generalizability in feature extraction. ResNet network models can be divided into ResNet-18, ResNet-34, ResNet-50, ResNet-101, and other structures. As different levels of features have different effects on the model, deep high-level features can help the model to classify, and low-level features can help the model regression. ResNet-101 can maintain high performance while ensuring a deep network layer. Therefore, to obtain a backbone feature extraction network more suitable for identifying various targets and improving the detection accuracy and speed of various small targets under complex backgrounds, the ResNet-101 network was selected to replace the VGG network. The network structure of ResNet-101 is shown in Fig. 3.

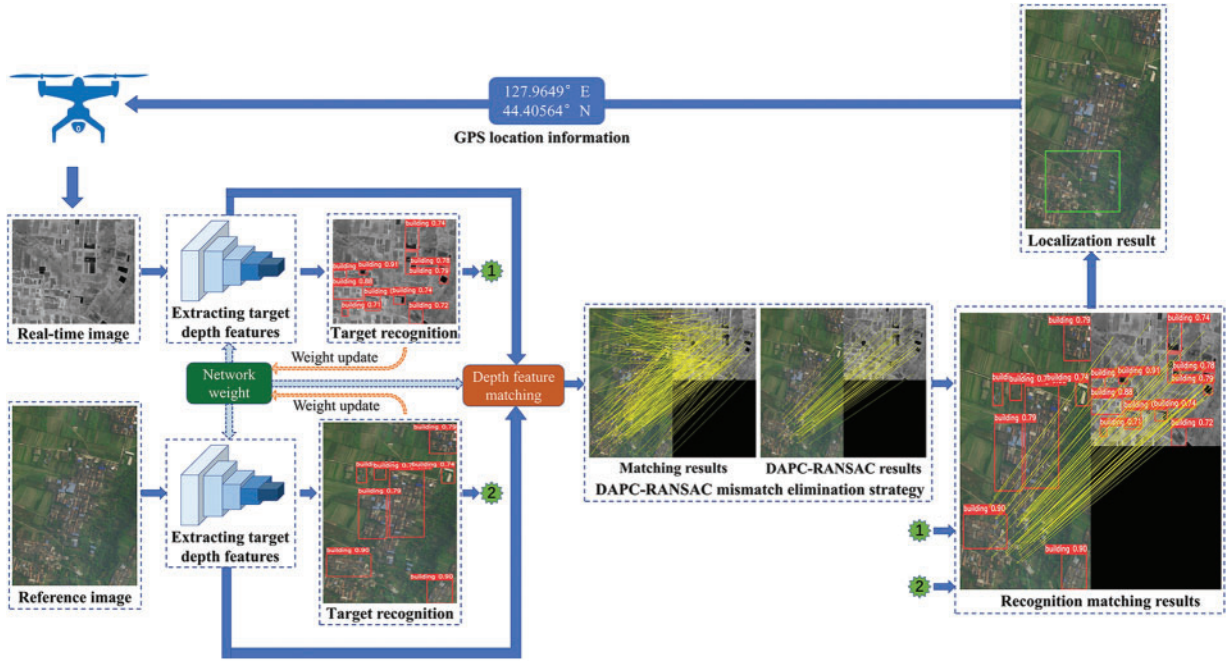


Figure 2: MCML algorithm flow chart

During the training of the M2Det algorithm, the deviation between the anchor frame and the rear frame will be large. To prevent gradient explosion and other situations, the Smooth L1 loss function was used in this paper. The calculation formula is as follows:

$$L_{SL1} = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & otherwise \end{cases} \tag{1}$$

where x is the difference between the prediction box and the real sample label. As the single-stage target recognition algorithm has the problem of imbalance between positive and negative samples in the training process, the improved M2Det target recognition algorithm in this paper added Focal Loss to calculate the classification loss during training. The Focal Loss calculation formula is as follows:

$$L_{fl} = \begin{cases} -\alpha\,(1-y')^{\gamma}\lg y', & y = 1 \\ -(1-\alpha)\,y'\lg(1-y'), & y = 0 \end{cases} \tag{2}$$

where y is the real sample label, y′ is the prediction output, α is the positive and negative sample weight, and γ is the weight of easy-to-classify samples and difficult-to-classify samples. Finally, the loss function used in this paper was the combination of Focal Loss and Smooth L1:

$$L = L_{fl} + L_{SL1} \tag{3}$$

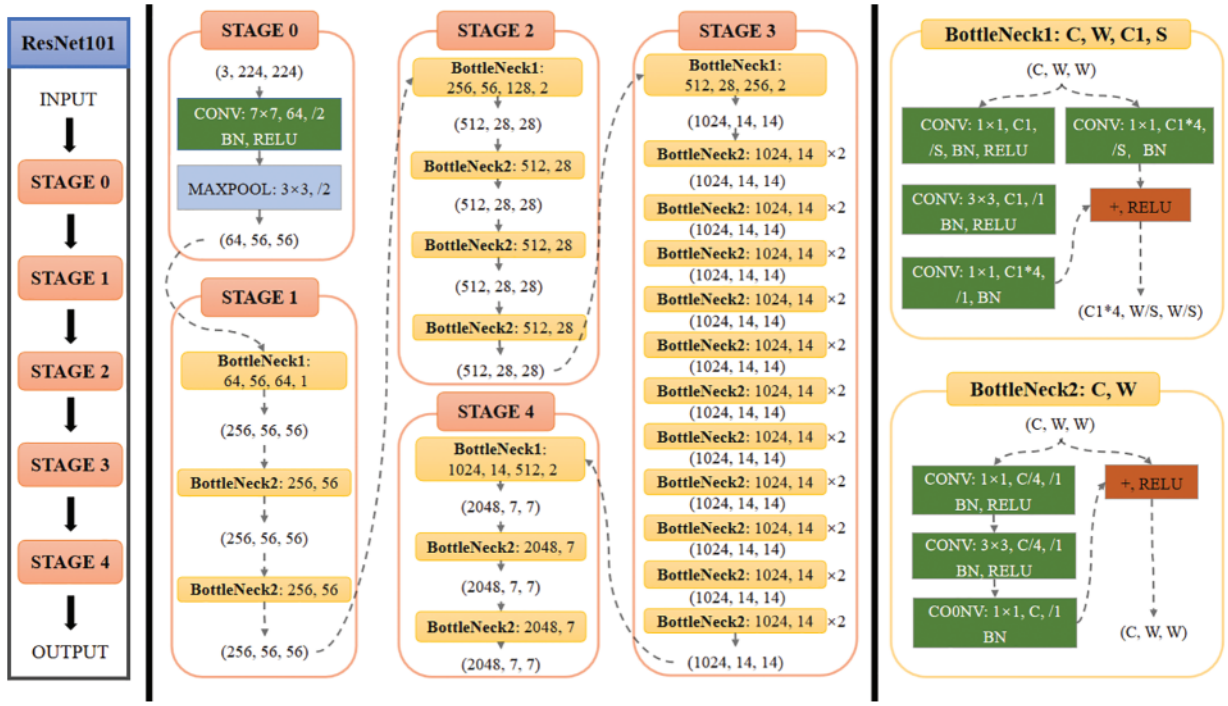


Figure 3: Network structure of ResNet-101

3.3 Depth Feature Matching Method

Depth feature matching is a method that uses the features extracted by the depth neural network to find the pixel correspondence between two images. It does not require external detection and feature description but directly calculates the correspondence between two images. The main idea of this paper was to use the ResNet-101 network, pre-trained by the M2Det target recognition algorithm, to extract features without any special training for feature matching. The classical ResNet-101 feature extraction is mainly used for classification tasks. Generally, the receptive field of the convolution network in the first few layers is very small, and the features obtained are generally corner, edge, and other features; however, the positioning accuracy is high. With the deepening of the network layer, the more abstract the extracted features are, the stronger the feature expression ability is, and the more complete the information is, the more resistant to the interference information between different source images; however, the positioning accuracy of features is poor. Therefore, to balance the contradiction between the abstraction of features and the positioning accuracy, this algorithm discarded the STAGE 2–4 features of ResNet-101 and selected the STAGE 1 output as the feature map for keypoint extraction. The feature map is thin-order toe output result of the original image after the multi-layer convolution and pooling of the ResNet-101 network. Since the resolution of the feature map will decline after each layer of the convolutional neural network is pooled, maintain the resolution of the feature map after

the pooling layer, this paper replaced the sliding step of the pooled layer window from 2 pixels to 1 pixel and replaced the maximum pooling with the average pooling. Assume that the input size $w \times h$ is the original image I, and set the feature map extracted by the network as a 3D tensor $F = \mathrm{F}(I)$, $F \in \mathrm{R}^{w \times h \times n}$, where the number of channels is $n = 512$. To filter out more significant feature points in the $\mathrm{R}^{w \times h \times n}$ feature space, the channel direction of the high-dimensional feature map and the maximum filtering strategy in the local plane is adopted. Set:

$$k = \arg\max D_{ij}^{t} \tag{4}$$

where D^k is the characteristic value of layer k, and $D^k \in \mathrm{R}^{w \times h}$, and D_{ij}^k is the characteristic values at the point (i,j) on the characteristic map. For point $P(i,j)$ to be selected, first select the channel k with the largest response value from the n channel characteristic graphs, obtain the characteristic graph on the corresponding channel is D^k, and finally verify whether D_{ij}^k is the local maximum. If these two conditions are met, the point to be selected $P(i,j)$ is a significant feature point. At the same time, 512-dimensional channel vector at (i,j) is extracted from feature map F and normalized to $L2$ normal form to obtain feature descriptor:

$$\widehat{d}_{ij} = d_{ij} / \left\| d_{ij} \right\|_2 \tag{5}$$

where $d_{ij} = F_{ij}, d \in \mathrm{R}^n$. Since the extreme points in discrete space are not real extreme points, to obtain more accurate positions of key points, the method of local interpolation of the feature map was used to obtain pixel-level positioning accuracy, and the descriptor was also obtained by bilinear interpolation in the neighborhood. Finally, $\widehat{d}_{ij}$ is a n dimensional vector obtained by interpolation, which can be used to match according to Euclidean space distance. Two methods of feature point matching are used in the Opencv3 open-source library: BF matcher matching and Flann-based matcher matching. Since the BF matcher matching method will try all the possibilities of matching points, it will find the global best matching point. However, in Flann-based matchers, the match is approximated. It will find the local best matching point, and the matching time is short. To improve the matching accuracy and obtain the global best matching point, a bidirectional BF matcher was selected for the matching operation in this paper. The matching algorithm flow is shown in Fig. 4.

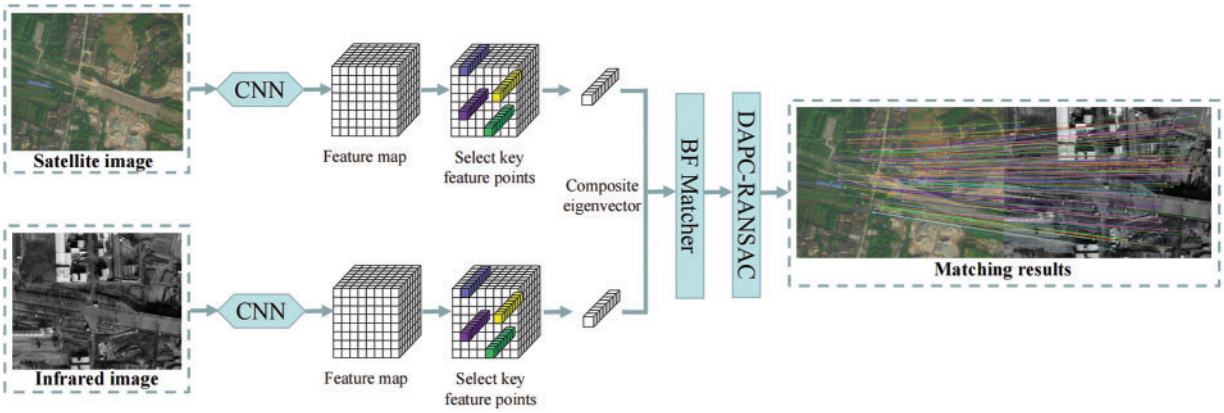


Figure 4: Depth feature matching algorithm

3.4 DAPC-RANSAC Mismatch Elimination Strategy

Due to the large difference between different source images, a large number of mismatches will inevitably occur in the process of image-matching. In this paper, a method combining dynamic adaptive

Euclidean distance constraints and random sampling consistency was used to eliminate mismatched point pairs. It is generally believed that for the j matching point pair in the matching point pair, the distance dis1 of the first matching point is smaller than the distance dis2 of the second matching point. It indicates that the matching point pair is better. Traditional algorithms use fixed threshold t to select candidate matching point pairs, that is, when $\text{dis1} < t \cdot \text{dis2}$ is satisfied, they are identified as matching point pairs. However, when the difference between different sensor imaging methods is large, it is difficult to distinguish the European space range when the depth feature is searched in European space. Therefore, each pair of images needs to manually adjust the threshold t to filter out the appropriate matching point pairs. To solve this problem, we automatically configured the corresponding threshold t according to the characteristics of the matching image. First, we counted the mean distance difference between the first matching point and the second matching point from the matching point pair:

$$\text{avgdis} = \left(\sum_{j=1}^{N} \text{dis}_j' - \text{dis}_j \right) / N \tag{6}$$

In the formula, N is the total matching point pair, dis is the distance between the first matching point, and dis′ is the distance between the second matching point; For each pair of matching points, the screening condition is that the first distance is less than the mean value of the second distance and distance difference, as follows:

$$\text{dis} = \text{dis}_j' - \text{avgdis} \tag{7}$$

The algorithm can adapt well to the differences between different source images and effectively filter the first round of feature-matching point pairs by obtaining the mean value of distance difference from the image-matching data as the criteria for discrimination and comparison. It provides a good initial value for the subsequent RANSAC algorithm, reduces the number of initial matching points, and improves the robustness and operation efficiency of the algorithm. The RANSAC algorithm is a widely used parameter estimation method. Using continuous iteration, it is able to find the optimal parameter model for a group of data sets with "matching points" and "mismatched points", and finally eliminate the mismatched points. Using bidirectional matching, this paper improved algorithm matching accuracy. The matching process was as follows:

(a) Images of image1 and image2 were read.

(b) The depth feature points of image1 and image2 were detected, and two sets of feature points, points1, and points2 were obtained, respectively.

(c) For each point i in points1, the corresponding point j in points2 was found.

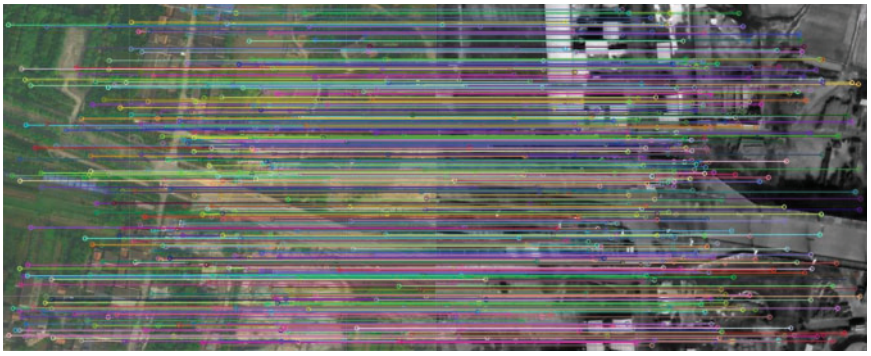
(d) For each point k in points2, the corresponding point l in points1 was found.

(e) If the matching point of point i was j and the matching point of point j was i, the matching was successful.

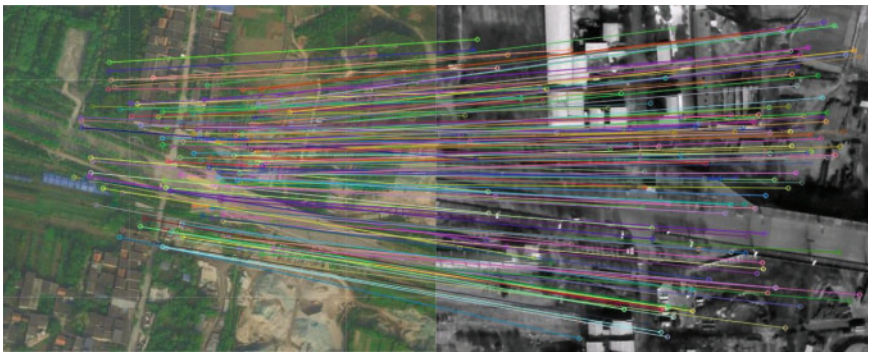
(f) The dynamic Euclidean distance between matching point pairs was calculated, and the first round of false matching point pair elimination was completed.

(g) The RANSAC algorithm was used to eliminate the second round of mismatched point pairs.

Fig. 5 shows the matching results of the two algorithms. Fig. 5 shows that the RANSAC algorithm solved the mismatch problem of the depth feature matching algorithm and obtained good matching results.



(a) Depth feature algorithm matching results.



(b) DAPC-RANSAC algorithm eliminates mismatch results.

Figure 5: Comparison of matching results of two algorithms

4 Experimental Results and Analysis

The proposed algorithm should be evaluated for feasibility and superiority, this paper used Opencv3 and MATLAB2016b in the University Release dataset and an S1000 six-rotor UAV aerial video for relevant experimental verification. The University Release datasets captured 1652 buildings from 72 universities, with a total of 50218 images, including three view images: ground streetscape, satellite, and drone perspectives. The operating system of the ground control station was Ubuntu 18.04, and the processor was an Inter (R) Core (TM) i7-11800U CPU@2.30 GHz 32 GB laptop. In this paper, nine classical matching algorithms were selected to compare the matching performance, and the UAV aerial video was used to analyze and verify the positioning effect of the algorithm.

4.1 Matching Efficiency Comparison

To verify the matching performance of the proposed algorithm, the visible image and satellite image, infrared image and satellite image, and infrared image and visible image in the multi-modal data set were selected for the matching and positioning experiments. The experimental scene had contrast, scale, brightness, blur, low resolution, and other scene changes. The experiment compared and analyzed SIFT, SURF, ORB, AKAZE, D2-Net, LoFTR, SuperPoint, Patch2Pix, AffNet+HardNet, R2D2, MCML, and other algorithms for feature matching. Scene A is the matching result between visible light image and satellite image, Scene B is the matching result between infrared image and satellite image, and Scene C is the matching result between infrared image and visible light image. The experimental results are shown in Fig. 6, and the performance comparison of matching methods

is shown in Tables 1–3. From the matching results, it can be seen that the MCML algorithm had a higher matching success rate, no false matching phenomenon that could effectively deal with the matching between multi-modal images, and had good adaptability to complex environments such as inter-image scale, blur, deformation, and low resolution. Due to the different application fields of image-matching methods, it is difficult to use a unified evaluation index to define the quality of image-matching results. In this paper, the performance of the algorithm was compared and analyzed from the matching positioning error. Matching positioning error refers to the proximity between the position information of the UAV and its real position according to the relative position relationship of matched feature points. This paper used the L2 distance to measure the positioning center error between real-time images and reference images. To evaluate the matching performance of the MCML algorithm in complex environments, Fig. 7 shows the comparison results of center position errors in three typical environments, and the legend shows the average values of center position errors of various algorithms in three environments. It can be seen from Figs. 6 and 7 that MCML had obvious advantages over the correct matching points of the traditional algorithms SIFT, SURF, ORB, and AKAZE. D2-Net, LoFTR, SuperPoint, Patch2Pix, R2D2, AffNet+HardNet, and other depth learning matching algorithms had many correct matching points; however, due to the existence of false matching, the positioning error was large. The matching efficiency comparison results from Tables 1 to 3 showed that the MCML algorithm had a good overall matching accuracy, good generalization and stable feature extraction ability for cross-modal image-matching, good robustness for contrast, scale, brightness, blur, deformation, and other changes, and a good positioning effect.

4.2 UAV Visual Localization Test

To verify whether the proposed algorithm can be implemented at night, we used a Pixhawk flight control board to independently build an S1000 six-rotor UAV target recognition and multi-modal scene matching fusion positioning test system. The complete UAV positioning and testing system included a ground station, dual optical pod, digital transmission equipment, wireless image transmission equipment, and an airborne computer for computing processing. The ground station was used to control UAV flights and monitor flight status. The real-time flight data of the UAV could be transmitted to the ground station in real-time through the data transmission equipment, and the input from the flight commander of the ground station could also be transmitted to the UAV. The data transmission equipment selected the 3DR radio data transmission radio V5 module, with a frequency of 915 MHz, a transmission power of 1000 mW, and a transmission distance of 5 km. The model of the dual optical pod used in this paper was TSHD10T3. The images collected by the pod were output by the HDMI interface, and the output frame rate was 60 FPS. The pod met the image acquisition requirements of target recognition and image-matching positioning and ensured real-time image recognition. In the process of target recognition and image-matching, the dual optical pod captured the real-time image and transmitted the image to the airborne computer for operation processing. In the airborne computer, the target recognition and image-matching positioning of image sequences were realized through the designed target recognition and a multi-modal scene-matching fusion positioning algorithm. Through the calculation of photographic geometry, the coordinates of the UAV and the target were obtained in the world coordinate system. Therefore, the position and posture of the UAV could be adjusted to achieve the visual positioning of the UAV. The configuration of UAV target recognition and multi-modal scene matching fusion positioning test system built in this paper is shown in Fig. 8.

Figure 6: Experimental comparison of various algorithms for test sequences

Table 1: Scene A: Performance comparison of image-matching methods

Methods	SIFT	SURF	ORB	AKAZE	D2-Net	LoFTR	SuperPoint	Patch2Pix	R2D2	AffNet+HardNet	MCML
Image A	1729	2405	500	342	3127	21	962	8	5000	1024	1388
Image B	1979	3148	500	719	3669	21	778	8	5000	1024	1406
Matching pair	148	187	171	99	508	21	266	8	108	27	96

Table 2: Scene B: Performance comparison of image-matching methods

Methods	SIFT	SURF	ORB	AKAZE	D2-Net	LoFTR	SuperPoint	Patch2Pix	R2D2	AffNet+HardNet	MCML
Image A	1686	2914	500	763	2736	35	726	16	5000	1024	855
Image B	829	1439	500	257	3501	35	1192	16	5000	1024	1315
Matching pair	204	262	169	171	597	35	288	16	112	6	161

Table 3: Scene C: Performance comparison of image-matching methods

Methods	SIFT	SURF	ORB	AKAZE	D2-Net	LoFTR	SuperPoint	Patch2Pix	R2D2	AffNet+HardNet	MCML
Image A	2915	2627	500	618	3462	72	688	16	5000	1024	978
Image B	578	1630	500	373	3614	72	719	16	5000	1024	1099
Matching pair	188	207	160	140	523	72	218	16	112	28	72

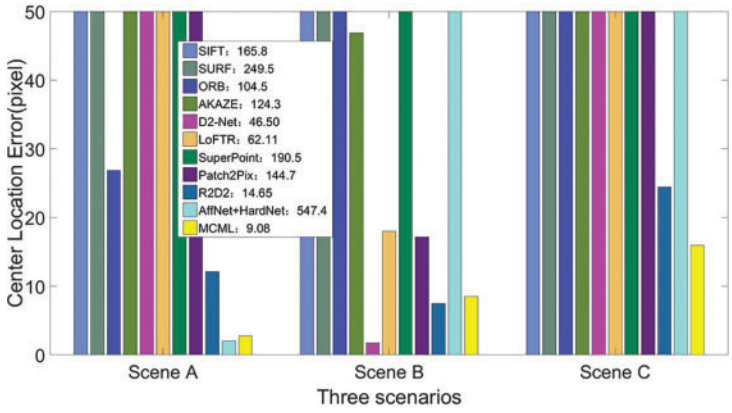


Figure 7: Comparison results of center position errors in three typical environments

The target recognition and multi-modal scene matching fusion location algorithm was based on a depth feature. The night infrared UAV location test scene is shown in Fig. 9, and the scene had unstructured environmental features. To improve the effectiveness of the proposed algorithm, the experimental environment used the UAV's forward and downward-looking flight angles of view. The resolution of the satellite reference image was set to 640 × 1064, the UAV's flight altitude was 495-500 m, and the flight distance was 2 km. The real-time image collected by the UAV was directly transmitted to the onboard computer through the pod network port for related operations. To improve the operation efficiency and save hardware computing resources, the aerial image was first preprocessed during the experiment, and the resolution of the infrared image collected by the pod was adjusted to 640 × 512. The real-time visualization results of the infrared image taken by the UAV are shown in Fig. 9. Fig. 10 represents the matching result of the UAV matching reference image and infrared image target recognition. It can be seen from Figs. 9 and 10 that the matching result of the

MCML algorithm was completely coincidental with the flight path, and the MCML algorithm had a good identification and matching result, which could output the corresponding center point of the real-time image in the reference image, to achieve the navigation and positioning of the UAV.

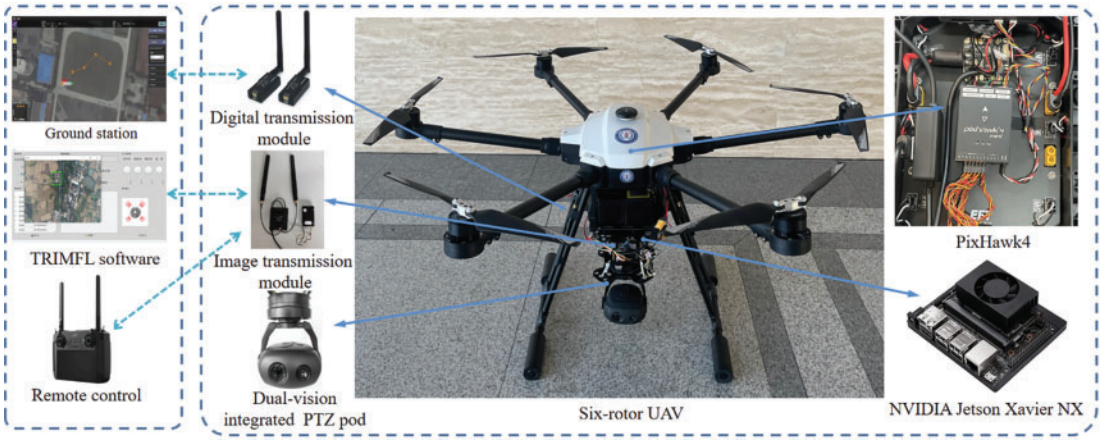


Figure 8: UAV visual localization test system

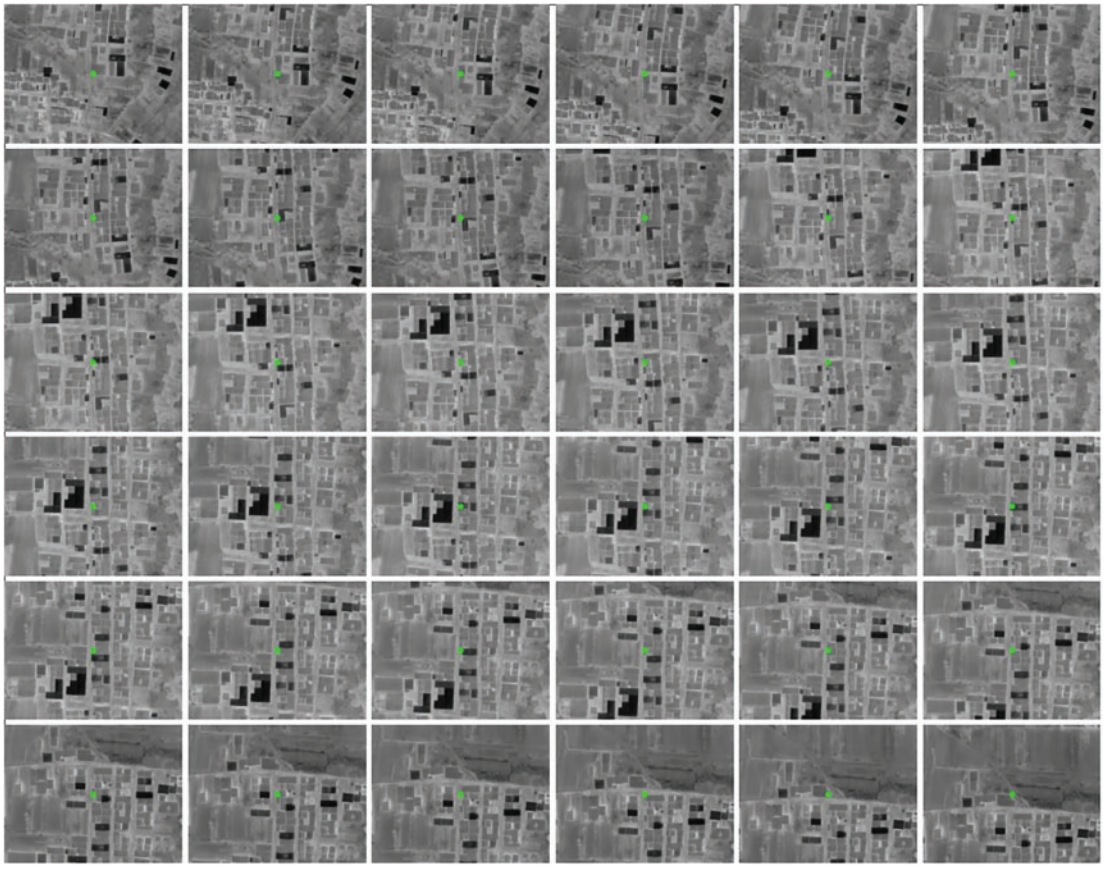


Figure 9: Real-time infrared image taken by UAV

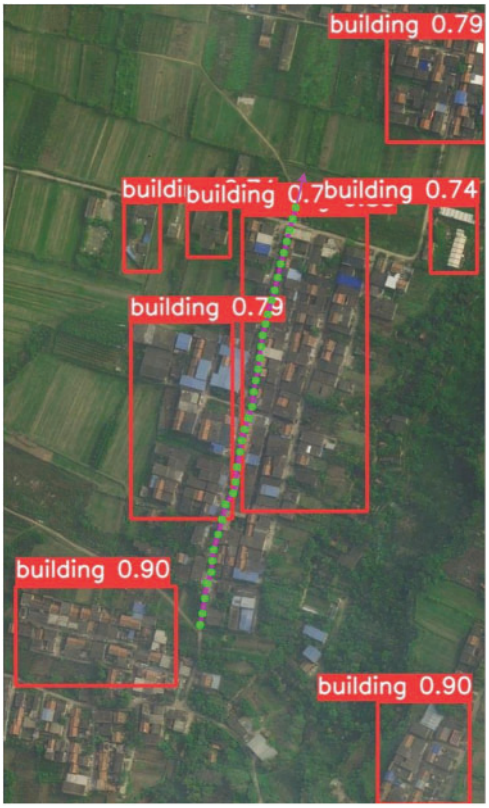


Figure 10: UAV matching reference map and target recognition matching results (●real-time map matching center point, ⟶ UAV flight path)

4.3 Discussion

Compared with the traditional image-matching methods, the MCML algorithm proposed in this paper greatly improved the matching accuracy and generalization of the algorithm; however, most of the depth learning algorithms need to rely on the powerful computing power of GPU. Recently, deep learning has been gaining popularity in the field of image-matching because the deep convolution neural network is trained using a large number of data, which extracts the deep features of the target and improves the matching effect. It is still difficult, however, to apply depth learning to image matching in practice: (i) image-matching algorithms need to have high real-time performance. In most depth learning image-matching algorithms, multi-layer convolutional neural networks are used to extract deep features from the image, thus improving the matching and positioning effect. Nevertheless, as convolution layers and the training network become more complex, higher requirements will be put forward for training samples and computations. (ii) The matching region is arbitrary, in this case, the image classification network may not be suitable for image matching and positioning because it was trained on a data set that had been used for image classification, which also presents a great test for the generalizability of the deep learning network. The algorithm proposed in this paper not only demonstrated the perceptibility of the target recognition algorithm to the environment, but also integrated the image-matching, and positioning function of depth features, which can effectively solve the problem of integration of target recognition, matching, and positioning, and make up for the shortcomings of a single algorithm. Although the operation speed is not as fast as some depth learning

image-matching methods, MCML is easy to implement and does not need more prior knowledge of the adaptive region. The image-matching effect is good, and it has certain engineering practical value.

5 Conclusion

The positioning function of UAVs is a challenging research topic but is essential for the autonomous navigation of UAVs. To solve the problem of real-time and robust matching when the UAV's heterogeneous images are quite different, this paper proposed a fusion localization algorithm of target recognition and multi-modal scene matching based on depth features. This algorithm extracts the depth features of reference images and real-time images by sensing the environmental information through the target recognition algorithm and uses the depth feature matching algorithm and the dynamic adaptive European distance random consistency error matching elimination strategy to complete target recognition and matching positioning tasks. The experimental results showed that the algorithm proposed in this paper has good adaptability to different flight environments and can complete matching and positioning tasks between infrared images and visible images, infrared images and satellite images, and visible images and satellite images. Compared with other matching algorithms, it has stronger robustness and higher matching accuracy. On the premise of ensuring real-time operation, it effectively improves the generalizability of the network, realizes the integrated design of target recognition and the matching positioning algorithm, and reduces the amount of computation. The recognition and matching performance are improved, especially when the illumination, scale, and imaging angle change greatly. Next, we plan to improve the algorithm's speed while ensuring that the matching positioning effect is maintained.

Acknowledgement: The authors are grateful to Zhengjie Zhu for her help with the preparation of figures in this paper.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 62276274, in part by the Natural Science Foundation of Shaanxi Province under Grant 2020JM-537, and in part by the Aeronautical Science Fund under Grant 201851U8012 (corresponding author: Xiaogang Yang).

Author Contributions: Conceptualization, X.Y., J.F. and R.L.; Methodology, J.F. and R.L.; Software, J.F.; Investigation, Q.L. and S.W.; Resources, Q.L.; Writing-original draft preparation, J.F. and R.L.; Writing-review and editing, X.Y., J.F. and S.W.; Visualization, J.F.; Supervision, J.F. and S.W.; Project administration, X.Y.; Funding acquisition, X.Y. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Wang and D. Liu, "Research on the applications of Russian precision strike weapons in the Russia-Ukraine conflict," *Tactical Missile Technology*, no. 3, pp. 107–115, 2022.

[2] F. Nex, C. Armenakis and M. Cramer, "UAV in the advent of the twenties: Where we stand and what is next," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 184, no. 1–4, pp. 215–242, 2022.
[3] Y. Guo, M. Wu and K. Tang, "Covert spoofing algorithm of UAV based on GPS/INS integrated navigation," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6557–6564, 2019.
[4] W. Youn, H. Ko and H. Choi, "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, 2020.
[5] G. M. Reich, M. Antoniou and C. Baker, "Memory-enhanced cognitive radar for autonomous navigation," *IET Radar Sonar and Navigation*, vol. 14, no. 9, pp. 1287–1296, 2020.
[6] C. Chi, X. Zhan and S. Wang, "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, 2020.
[7] 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. https://doi.org/10.1007/s00138-019-01046-4
[8] S. Chen, H. Chen and W. Zhou, "End-to-end UAV simulation for visual SLAM and navigation," *Aerospace*, vol. 9, no. 2, pp. 1–16, 2022.
[9] 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.
[10] S. H. Choi and C. G. Park, "Image-based Monte-Carlo localization with information allocation logic to mitigate shadow effect," *IEEE Access*, vol. 8, pp. 213447–213459, 2020. https://doi.org/10.1109/ACCESS.2020.3039413
[11] M. H. Mughal, M. J. Khokhar and M. Shahzad, "Assisting UAV localization via deep contextual image-matching," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 140, pp. 2445–2457, 2021. https://doi.org/10.1109/JSTARS.2021.3054832
[12] L. Ding, J. Zhou and L. Meng, "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, 2020.
[13] X. Jiang, J. Ma and G. Xiao, "A review of multimodal image-matching: Methods and applications," *Information Fusion*, vol. 73, no. 11, pp. 22–71, 2021. https://doi.org/10.1016/j.inffus.2021.02.012
[14] 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. https://doi.org/10.1109/JSTARS.2020.2982929
[15] 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, 2020.
[16] 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.
[17] K. Yu, C. Xu, J. Ma and B. Fang, "Automatic matching of multimodal remote sensing images via learned unstructured road feature," *Remote Sensing*, vol. 14, no. 18, pp. 1–19, 2022.
[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[19] E. Rublee, V. Rabaud, K. Konolige and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, pp. 2564–2571, 2011.
[20] Y. Niu, M. Chen and H. Zhang, "Fast scene matching method based on scale invariant feature transform," *Journal of Electronics & Information Technology*, vol. 41, no. 3, pp. 626–631, 2019.
[21] C. Yuan and X. Sun, "Fingerprint liveness detection using histogram of oriented gradient based texture feature," *Journal of Internet Technology*, vol. 19, no. 5, pp. 1499–1507, 2018.

[22] 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. 22, no. 78, pp. 31893–31924, 2019.
[23] Y. Zhang, J. Song and Y. Ding, "FSD-BRIEF: A distorted BRIEF descriptor for fisheye image based on spherical perspective model," *Sensors*, vol. 21, no. 5, pp. 1–26, 2021.
[24] J. Li, Q. Hu and M. Ai, "RIFT: Multi-modal image-matching based on radiation-variation insensitive feature transform," *IEEE Transactions on Image Processing*, vol. 29, pp. 3296–3310, 2020. https://doi.org/10.1109/TIP.2019.2959244
[25] J. Li, W. Xu, P. Shi and Y. Zhang, "LNIFT: Locally normalized image for rotation invariant multimodal feature matching," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022. https://doi.org/10.1109/TGRS.2022.3165940
[26] C. Dai, C. Peng and M. Chen, "Selective transfer cycle GAN for unsupervised person re-identification," *Multimedia Tools and Applications*, vol. 79, no. 18, pp. 12597–12613, 2020.
[27] Y. Luo, D. Pi and Y. Pan, "ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light," *Expert Systems with Application*, vol. 191, no. 3, pp. 1–13, 2022. https://doi.org/10.1016/j.eswa.2021.116269
[28] Z. Zuo, L. Zhao and A. Li, "Dual distribution matching GAN," *Neurocomputing*, vol. 478, pp. 37–48, 2022. https://doi.org/10.1016/j.neucom.2021.12.095
[29] G. Chang, W. Wang and S. Hu, "Match ACNN: A multi-granularity deep matching model," *Neural Processing Letters*, pp. 1–20, 2022. https://doi.org/10.1007/s11063-022-11047-6
[30] X. Zan, X. Zhang and Z. Xing, "Automatic detection of maize tassels from UAV images by combining random forest classifier and VGG16," *Remote Sensing*, vol. 12, no. 18, pp. 1–17, 2020.
[31] M. Bansal, M. Kumar and M. Sachdeva, "Transfer learning for image classification using VGG19: Caltech-101 image data set," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, no. 17, pp. 1–12, 2021.
[32] K. He, X. Zhang and S. Ren, "Deep residual learning for image recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, NV, USA, pp. 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90
[33] O. Zhang, "A novel ResNet101 model based on dense dilated convolution for image classification," *SN Applied Sciences*, vol. 4, no. 1, pp. 1–13, 2022. https://doi.org/10.1007/s42452-021-04897-7
[34] G. Li, M. Zhang and J. Li, "Efficient densely connected convolutional neural networks," *Pattern Recognition*, vol. 109, pp. 1–9, 2020. https://doi.org/10.1016/j.patcog.2020.107610
[35] F. Fooladgar and S. Kasaei, "Lightweight residual densely connected convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 35, pp. 25571–25588, 2020.
[36] X. Zhang, R. Jiang and T. Wang, "Single image dehazing via dual-path recurrent network," *IEEE Transactions on Image Processing*, vol. 30, pp. 5211–5222, 2021. https://doi.org/10.1109/TIP.2021.3078319
[37] J. Wang, Z. Shao and X. Huang, "A dual-path fusion network for pan-sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021. https://doi.org/10.1109/TGRS.2021.3090585
[38] H. S. Kim, K. Y. Yoo and L. H. Kim, "Improved performance of image semantic segmentation using NASNet," *Korean Chemical Engineering Research*, vol. 57, no. 2, pp. 274–282, 2019.
[39] X. Han, T. Leung and Y. Jia, "MatchNet: Unifying feature and metric learning for patch-based matching," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 3279–3286, 2015.
[40] H. Deng, T. Birdal and S. Ilic, "PPFNet: Global context aware local features for robust 3D point matching," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 1–11, 2018. https://doi.org/10.1109/CVPR.2018.00028
[41] J. Sun, Z. Shen and Y. Wang, "LoFTR: Detector-free local feature matching with transformers," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8918–8927, 2021.

[42] E. Simo-Serra, E. Trulls and L. Ferraz, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 118–126, 2015.
[43] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 4353–4361, 2015.
[44] T. Chen and J. Guo, "Visible and infrared image-matching method based on generative adversarial model," *Journal of Zhejiang University (Engineering Science)*, vol. 56, no. 1, pp. 63–74, 2022.
[45] H. Zhang, W. Ni and W. Yan, "Registration of multi-modal remote sensing image based on deep fully convolutional neural network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3028–3042, 2019.
[46] M. Dusmanu, I. Rocco and T. Pajdla, "D2-Net: A trainable cnn for joint description and detection of local features," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, pp. 8092–8101, 2019. https://doi.org/10.1109/CVPR.2019.00828
[47] J. Revaud, C. D. Souza and M. Humenberger, "R2D2: Reliable and repeatable detectors and descriptors for joint sparse keypoint detection and local feature extraction," *Advances in Neural Information Processing Systems*, pp. 1–13, 2019. https://doi.org/10.48550/arXiv.1906.06195
[48] P. E. Sarlin, D. Detone and T. Malisiewicz, "SuperGlue: Learning feature matching with graph neural networks," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 4937–4946, 2020. https://doi.org/10.1109/CVPR42600.2020.00499
[49] A. B. Laguna, E. Riba and D. Ponsa, "Key.Net: Keypoint detection by handcrafted and learned CNN filters," in *Proc. of IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, Seoul, Korea, pp. 5835–5843, 2019. https://doi.org/10.1109/ICCV.2019.00593
[50] V. Balntas, E. Johns and L. Tang, "PN-Net: Conjoined triple deep network for learning local image descriptors," arXiv preprint arXiv: 1601.05030, 2016.
[51] Y. Ye, L. Bruzzone and J. Shan, "Fast and robust matching for multi-modal remote sensing image registration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9059–9070, 2019.
[52] Z. Li, H. Zhang and Y. Huang, "A rotation-invariant optical and SAR image registration algorithm based on deep and gaussian features," *Remote Sensing*, vol. 13, no. 13, pp. 1–24, 2021.
[53] D. Mishkin, F. Radenovic and J. Matas, "Repeatability is not enough: Learning affine regions via discriminability," in *Proc. of European Conf. on Computer Vision (ECCV)*, Munich, Germany, pp. 1–18, 2018. https://doi.org/10.1007/978-3-030-01240-3_18
[54] J. Redmon, S. Divvala and R. Girshick, "You only look once: Unified, real-time object detection," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 429–442, 2016. https://doi.org/10.1109/CVPR.2016.91
[55] Y. Xue, Z. Ju and Y. Li, "MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection," *Infrared Physics & Technology*, vol. 118, pp. 1–14, 2021. https://doi.org/10.1016/j.infrared.2021.103906
[56] X. Li, "A real-time detection algorithm for kiwifruit defects based on YOLOv5," *Electronics*, vol. 10, no. 14, pp. 1–13, 2021.
[57] X. Ding, Q. Li and Y. Cheng, "Local keypoint-based faster R-CNN," *Applied Intelligence*, vol. 50, no. 10, pp. 3007–3022, 2020.
[58] J. Qin, Y. Zhang and H. Zhou, "Protein crystal instance segmentation based on mask R-CNN," *Crystals*, vol. 11, no. 2, pp. 1–8, 2021.
[59] M. Wu, H. Yue and J. Wang, "Object detection based on RGC mask R-CNN," *IET Image Processing*, vol. 14, no. 8, pp. 1502–1508, 2020.
[60] D. Jia, J. Zhou and C. Zhang, "Detection of cervical cells based on improved SSD network," *Multimedia Tools and Applications*, vol. 81, no. 10, pp. 13371–13387, 2021.

[61] 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. https://doi.org/10.1007/s11227-020-03578-3

[62] X. Zhang, H. Xie and Y. Zhao, "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. https://doi.org/10.1007/s11554-021-01108-9

[63] S. J. Hong, W. K. Baek and H. S. Jung, "Ship detection from X-band SAR images using M2Det deep learning model," *Applied Sciences*, vol. 10, no. 21, pp. 1–19, 2020.

[64] K. Zhang and L. Zhang, "Multi-scale detection for X-ray prohibited items in complex background," *Laser & Optoelectronics Progress*, vol. 58, no. 22, pp. 1–11, 2021.

[65] Q. Zhao, T. Sheng and Y. Wang, "M2Det: A single-shot object detector based on multi-level feature pyramid network," in *Proc. of AAAI Conf. on Artificial Intelligence*, Hawaii, USA, vol. 33, pp. 9259–9266, 2019.