IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 September 2022, accepted 19 October 2022, date of publication 26 October 2022, date of current version 3 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3217235

RESEARCH ARTICLE

A Fast Template Matching Scheme of Visible and Infrared Image Under Occluded Scenarios

LICHUN MEI[1], HUAIYE WANG[2], CAIYUN WANG[3], YUANFU ZHAO[4], (Senior Member, IEEE), JUN ZHANG[2], AND XIAOXIA ZHAO[2]

[1]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[2]Beijing Space Feiteng Equipment Technology Company Ltd., Beijing 100094, China
[3]College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
[4]Beijing Microelectronic Technology Institute, Beijing 100094, China

Corresponding author: Lichun Mei (mlchun@nuaa.edu.cn)

ABSTRACT A fast and robust template matching scheme, called Matching by Slice Transform Matrix Mapping (MSTMM), is proposed for the matching difficulty under occluded scenarios caused by nonlinear intensity differences and structure differences between visible and thermal infrared images. The first step in the MSTMM scheme was to extract information about the distribution of pixels with the same gray level through the developed Expanded Slice Transform with Adaptive Gray Level (EST-AGL). After completing the construction of the EST-AGL matrix for all image patches, different EST-AGL matrices were mapped to different integers or floats through the traditional special integer mapping mechanism or the neural network mapping mechanism. Finally, template matching between visible and thermal infrared images was achieved by evaluating the similarity of correlation mapping surface images through the Normalized Cross Correlation (NCC) algorithm. The proposed EST-AGL method can overcome the nonlinear intensity differences between visible and thermal infrared images by extracting the structural features of the image. The mapping mechanism of the MSTMM scheme can reduce the structural differences between the normal template image and the query image under an occluded scenario by increasing the similarity between the normal image patches and the image patches with occlusion. The proper mapping mechanism ensures the high performance of the MSTMM scheme by using only the simple NCC algorithm instead of other time-consuming anti-occlusion dense feature algorithms in the similarity evaluation stage. The three main experimental results of the MSTMM scheme are as follows: (1) the scheme of MSTMM can achieve template matching in only 0.015 seconds when a 64×64 template image slides on a 256×256 query image on a hardware platform with limited resources; (2) the matching success rate of the MSTMM scheme can reach up to 75% among 2107 experimental samples; and (3) the neural network training in the neural network mapping mechanism only takes at least 104.4 seconds on the CPU.

INDEX TERMS Template matching, multimodal image, heterogeneous image, multisource image, visible and infrared image, image matching, neural network.

I. INTRODUCTION

With the development of science and technology, the visible imaging system has spread to almost all aspects of society and life. The infrared thermal imaging system is different from the visible imaging system in the imaging principle and application field [1]. Infrared thermal imaging system is mainly passive infrared imaging, which uses real-time acquisition of the difference in infrared thermal radiation intensity of different parts of natural objects to form images [2], [3], [4], [5], [6], so it is widely used in electric power prevention and detection [7], industrial temperature measurement [8], medical quarantine [9], auto-auxiliary driving [10], fire search and rescue [11], security monitoring [12] and other fields. In these special fields, the addition of an infrared thermal imaging system avoids the defects of the original visible imaging system

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa Rahimi Azghadi.

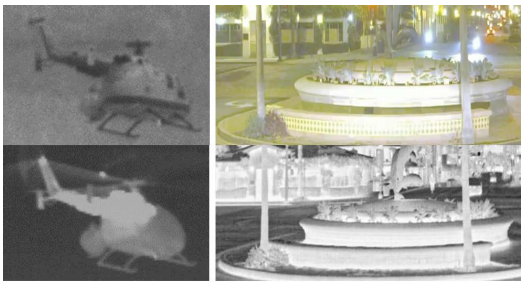
and promotes the rapid development of related fields. The problem that follows is how to make full use of the different images built by these different imaging systems to promote technological development in related fields. Template matching between visible and thermal infrared images is one of the very important and fundamental problems.

Existing template matching methods can be roughly divided into the traditional method and the deep learning method based on neural networks. Traditional methods usually measure the similarity between two images through the statistical difference of grayscale information or the distance between dense feature vectors [13], [14], [15], [16], such as Normalized Cross Correlation (NCC) [17], [18], [19], Sum of Squared Difference (SSD) [20], [21], Mutual Information (MI) [22], Local Self-Similarity (LSS) [23], [24], etc. The pixel-level grayscale information statistics and complex dense feature extraction lead to the inefficiency of this type of algorithms in the early stage. However, the introduction of Fast Fourier Transform (FFT) through the convolution theorem [25] has brought a significant improvement to the efficiency of some pixel-level algorithms, and more ingenious features in [13] can also reduce the difficulty of dense feature extraction. Deep learning methods based on the neural network usually use deep features extracted by Convolutional Neural Networks (CNN) to perform template matching [26], such as Quality Aware Template Matching (QATM) [27], Bottom-Up Pattern Matching (BUPM) [28], etc. At present, since the template matching problem focuses on measuring the similarity between the template image and candidate windows of the query image, it is difficult to directly perform template matching through CNN deep learning algorithms based on classification problems. Currently, the traditional method can be executed on almost any hardware platform, while the neural network model of the deep learning method usually requires larger memory for storage, and the training process of the model has high requirements on the hardware platform, so it is difficult to realize online training.

The visible image is constructed by a visible light sensor by capturing different reflections of light from different object surfaces, while the thermal infrared image is constructed by an infrared sensor by capturing the difference in infrared thermal radiation intensity of different parts of natural objects [2], [3], [4], [5], [6]. The large differences between visible and thermal infrared images pose a very significant challenge for template matching schemes, whereas most existing template matching methods rely on linear, monotonic, or functionally constrained matching rules [29], [30], [31]. Special algorithms are required for template matching between heterogeneous images. At present, most of the traditional heterologous image template matching algorithms evaluate the similarity of two images through the distance between dense features, such as Histogram of Orientated Phase Congruency (HOPC) [32], Structure Tensor Voting and Orientation (STVO) in [13], Local Central-Tendency Similarity (LCTS) in [33], etc. Currently, CNN-based deep learning algorithms are rarely used in template matching of heterologous images, and apart from the problems they face in template matching of homologous images, a large number of aligned heterologous image datasets are currently scarce. Therefore, it is necessary to construct an aligned heterologous image dataset, such as Normalized Cross Correlation Network (NCCNet) [34], Matching RGB and Infrared images (M-RGBIR) in [35], etc. Note that since the original papers corresponding to LCTS, STVO, and M-RGBIR do not give specific algorithm names, this paper uses the abbreviated names LCTS, STVO, and M-RGBIR to represent the algorithm names in the corresponding papers.

Currently, there is an important branch in the field of deep learning called image translation [36], [37], [38]. We can use image translation to convert images of different modalities into images of the same modality so that we can use the existing template matching algorithm for homologous images to achieve template matching between heterologous images. This is also an important source of ideas for the algorithm in this paper. Currently, most image translation algorithms are built on Generative Adversarial Network (GAN) [39]. The generative adversarial network received extensive research and high attention from academia and industry after Pix2pix [38]. With the indepth study of GAN and Pix2pix, Multimodal Unsupervised Image-to-image Translation (MUNIT) [40] further supports the translation of multimodal images.

In a visible image, the grayscale, form, and texture an object presents are determined by the object's ability to reflect light, while in a thermal infrared image, they are determined by how much thermal radiation is captured in different parts of natural objects [41]. Due to different imaging mechanisms, the normal object usually shows rich textures on visible images, while spots of different gray values appear on the thermal infrared image, as shown in Fig. 1. In order to match the texture-rich objects in the visible image, this study treats spots of different gray values in the thermal infrared image as occlusions for texture-rich objects. At present, template matching algorithms for complex scenes such as occlusion and deformation are mainly limited to matching between homologous images, such as Best Buddies Similarity (BBS) [42], Deformable Diversity Similarity (DDIS) [43], Occlusion Aware Template Matching (OATM) [44], Siamese Network in [45], Structure Tensor Voting and Orientation (STVO) in [13], etc. In this study, we simulate the different sizes of spots on the thermal infrared image to study the impact of different occlusion degrees on the matching algorithm. In addition, considering the different requirements for real-time and robustness of different hardware platforms, the algorithmic architecture needs to be flexible enough to meet different requirements. Motivated by the limitations of current methods and practical requirements, we propose Matching by Slice Transform Matrix Mapping (MSTMM), a template matching scheme of visible and thermal infrared images based on Expanded Slice Transform with Adaptive Gray Level (EST-AGL), which attempts to combine the advantages of traditional methods and neural network



(a) (b)

FIGURE 1. The examples of visible imaging and infrared imaging results: (a) image pair from [46]; (b) image pair from [47]. The image pairs of (a) and (b) are visible image on the top and thermal infrared image on the bottom.

methods to solve the problems encountered in current template matching for visible and thermal infrared images.

Different from the mapping method in Matching by Tone Mapping (MTM) [48], the MSTMM mapping method proposed in this paper is more similar to the GAN-based image translation method, i.e., one image is mapped to another image according to certain rules. The MSTMM architecture proposed in this paper is shown in Fig. 2. The processes of the MSTMM template matching scheme are as follows: first, input the visible image or the image patch cropped from the visible image as the template image, and input the thermal infrared image as the query image; then, using the traditional special integer mapping mechanism or neural network mapping mechanism proposed in this paper, the template image and query image of different modalities are converted into corresponding mapping images of the same modality; finally, template matching is realized by using the NCC template matching algorithm. The scheme of MSTMM can use three flexible mapping implementation methods to meet the needs of different hardware platforms, which are the traditional special integer mapping method, the offline neural network training method, and the online neural network training method. Note that the detailed parameters of the spots with different gray values in the thermal infrared image in Fig. 1, please refer to the experimental part of Section IV.

The proposed work has three major contributions. First, we developed an Expanded Slice Transform with Adaptive Gray Level (EST-AGL) to extract information about the distribution of pixels with the same gray level. The Slice Transform (SLT) matrix [48] transforms a grayscale image into a distribution matrix of pixel points within a defined gray value interval size that encompasses the whole image. Since the distribution matrix not only is unaffected by the specific gray values but can also reflect the image's structural information, it can effectively mitigate the negative effects of nonlinear intensity differences between heterogeneous images on subsequent processing steps. However, the dimension of the SLT matrix is determined by the size of the interval and the maximum and minimum gray values on the whole image, which restricts its applicability. The dimension of the EST-AGL matrix we developed is only related to the number of pixel points on the whole image so that each pixel point can get the distribution matrix of the corresponding position on the whole image. The method of EST-AGL can overcome the nonlinear intensity differences between visible and thermal infrared images by extracting the structural features of the image.

Second, to meet the requirements of different hardware platforms, we propose two mapping mechanisms to map input images of different modalities to the mapped correlation surface images of the same modality, i.e., the special integer mapping mechanism and the neural network mapping mechanism. The two mapping mechanisms can reduce the structural differences between the normal template image and the query image under an occluded scenario by increasing the similarity between the normal image patches and the image patches with occlusion. The special integer mapping mechanism can be used as an independent traditional mapping method to achieve template matching. The neural network mapping mechanism can achieve the purpose of improving the robustness of the MSTMM scheme. The offline network training method can improve the robustness of the MSTMM scheme without increasing hardware resource requirements, and the online network training method can improve the robustness of the MSTMM scheme in complex and varied heterogeneous video scenes. The two proper mapping mechanisms ensure the high performance of the MSTMM scheme by using only the simple NCC algorithm instead of other time-consuming anti-occlusion dense feature algorithms in the similarity evaluation stage.

Finally, the usage of a minimalist fully connected feedforward neural network (FNN) rather than the popular convolutional neural network (CNN) brings the possibility of online training. For the minimalist FNN, a variety of online training solutions are developed to satisfy the matching performance under different hardware resources. In addition, since the network model of Matching by Slice Transform Matrix Mapping of Network Mapping (MSTMM-NM) mechanism proposed in this paper has only a limited number of weights, and the weights are only related to the size of the mapped value, the trained model can directly extract these weights into a ''.txt'' file, and then use these weights in the imported ''.txt'' file for template matching. The advantage of this is that the trained model can handle input images of arbitrary size, avoiding the problem of most deep learning algorithms accepting only fixed-size input images for inference.

The remainder of this paper is the following: In Section II the related works are introduced. The proposed template matching scheme of visible and thermal infrared images is presented in Section III. In Section IV some algorithm configuration parameters and comparative experiments and training solutions are conducted to prove the superiority of the scheme proposed in this paper. Finally, Section V presents the conclusions, limitations, and some future works.

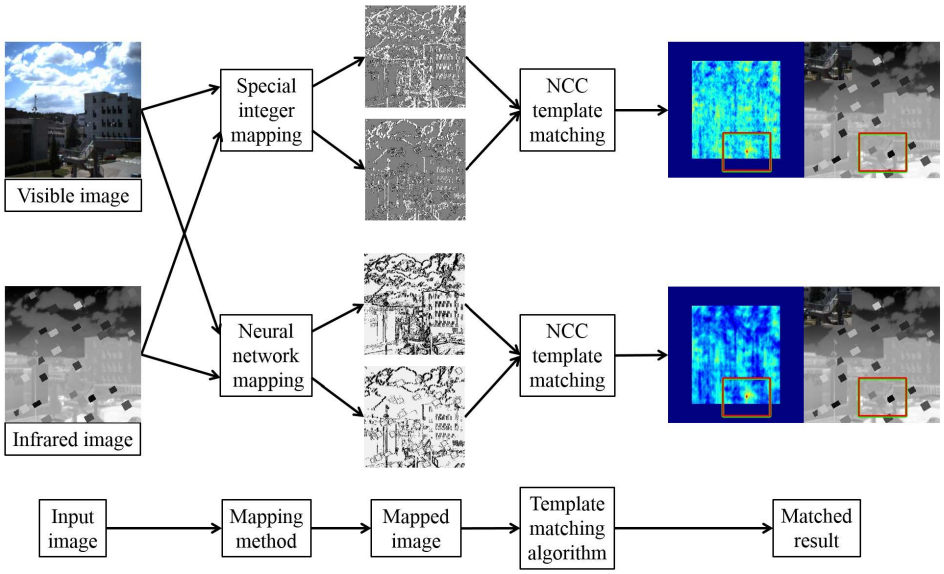


FIGURE 2. The MSTMM architecture and template matching process.

II. RELATED WORKS

The Normalized Cross Correlation (NCC) is a relatively common cross correlation calculation method in template matching, which achieves the similarity measure between two vectors, windows, or samples by describing the correlation between them. The algorithm of NCC is unaffected by variations in linear intensity, in general, which can perform well for monotonic nonlinear intensity mapping [13]. However, when the monotonicity mapping is disrupted, the NCC algorithm will fail [13].

The Sum of Squared Difference (SSD) [20], [21] is quite sensitive to radiometric differences because they directly compute the differences in the intensity between the images. Therefore, the algorithm of SSD is not suitable for matching the visual and thermal infrared images. However, the algorithm of SSD is often applied to multisource image matching as an auxiliary algorithm of the Local Self-Similarity (LSS) [49] class algorithm. The algorithm of LCTS [33] is a heterologous image template matching algorithm based on the LSS class algorithm. The algorithm of LCTS calculates similarity by using local centraltendency and gaussian-weighting function and then measures similarity between the image pairs by using Best-Buddies Direction Pairs (BBDP) algorithm based on Best-Buddies Similarity (BBS) [42]. The computational efficiency of the LSS class algorithm is low since it belongs to a dense feature descriptor algorithm and an SSD algorithm is set inside it. The computational efficiency of the LCTS algorithm combined with the inefficient BBDP based on BBS is lower. And because the LCTS algorithm needs to extract a large area around the pixel to calculate the local descriptor, the matching result is not good on the area of the query image border.

At present, most multisource image template matching algorithms achieve template matching by extracting structural features from the grayscale image and then evaluating the similarity between dense structural features [13], [14], [15], [16]. Recently, Lu et al. [13] proposed a template matching algorithm based on Structure Tensor Voting and Orientation (STVO), which is a dense feature descriptor algorithm but has a much higher computing efficiency. The descriptor of STVO is built on a dense structure tensor that successfully captures the structural features of noise-degraded images in complex scenes. However, after experimental comparison, its performance and computational efficiency in visible and thermal infrared image template matching still need to be improved.

The Matching by Tone Mapping (MTM) [48] can measure similarity between heterogeneous images under monotonic and nonmonotonic nonlinear intensity mapping. The algorithm of MTM is a generalization of the NCC algorithm under nonmonotonic nonlinear intensity mapping, and it reduces to the NCC algorithm when the mapping is restricted to being monotonic. The advantage of the MTM algorithm is that it is very fast in execution, however, after experimental comparison, its performance in visible and thermal infrared images template matching still has space for improvement.

To build a fast template matching scheme, we focus on the MTM algorithm and build a fast template matching scheme for multisource images. Although the algorithm of MTM essentially realizes the similarity measurement between the template and the candidate window by calculating the similarity of the grayscale information, we discovered that the SLT matrix used in the MTM algorithm can be improved to extract information about the distribution of pixels with the same gray level. The distribution information of image pixels can also be used as a kind of special structural information to overcome nonlinear intensity differences between

heterogeneous images, and finally realize the similarity measure between visible and thermal infrared images.

The algorithms mentioned above are traditional algorithms. With the development of science and technology, the deep learning algorithm based on the neural network has been applied in all walks of life and achieved good results, especially in the field of computer vision.

The CNN-based template matching network Normalized Cross Correlation Network (NCCNet) [34] maximizes the contrast between true and false matching NCC values by transforming image features using a trained Siamese convolutional network, thereby improving the robustness of the algorithm. The algorithm of NCCNet is a weakly supervised learning algorithm, which, unlike fully supervised metric learning methods, improves the computational process of ordinary NCC template matching algorithms without receiving true matching positions during training. The algorithm of NCCNet is achieved by maximizing the NCC maximum and submaximal values. However, the algorithm of NCC has poor matching performance between visible and infrared images with large nonlinear differences, and the NCC maximum point is not the ground truth point, so it is meaningless to maximize the NCC maximum and submaximum values.

The algorithm of Matching RGB (red, green, and blue) and Infrared images (M-RGBIR) in [35] is a deep learning-based matching algorithm between RGB and infrared images. The algorithm uses a densely connected CNN to extract common features in RGB and infrared image pairs, enabling matching between heterogeneous images. The densely connected CNN in M-RGBIR can fully utilize low-level features and augmented cross-entropy loss to avoid model overfitting. The network in M-RGBIR takes as input the RGB and infrared images of the band concatenation and outputs a similarity score of the RGB and infrared image pairs. For a given template, the algorithm of M-RGBIR uses a sliding window on the query image to slide pixel-by-pixel to measure the similarity between the template and the subwindow in the query image, and the position of the subwindow with the highest score is the position of the matching template. The algorithm of M-RGBIR has good generalization ability, but since each subwindow has to go through a complex dense network to achieve similarity measurement, this algorithm is inefficient in sliding windows. After experimental comparison, it takes about 1528.46 seconds for M-RGBIR to implement a complete sliding window process on a 256 × 256 query image with a 64 × 64 template image.

The GAN-based Multimodal Unsupervised Image-to-image Translation (MUNIT) [40] algorithm is a deep learning algorithm proposed in 2018 to support multimodal image translation. The algorithm of MUNIT assumes that image representations can be decomposed into a domain-invariant content code and a style code that captures domain-specific properties. To convert an image to another domain, the content code of the original image is recombined with a style code randomly selected from the target domain. The algorithm of MUNIT can translate images of one modality into images of another modality by decomposing the content and style spaces. However, the model of MUNIT is relatively difficult to train and requires large amounts of data to train. After experimental comparison, the algorithm of MUNIT does not work well to translate visible images into thermal infrared images with large nonlinear differences, and this translation process will cause the input image to lose a lot of detail, which greatly affects the performance of subsequent homologous image template matching algorithm.

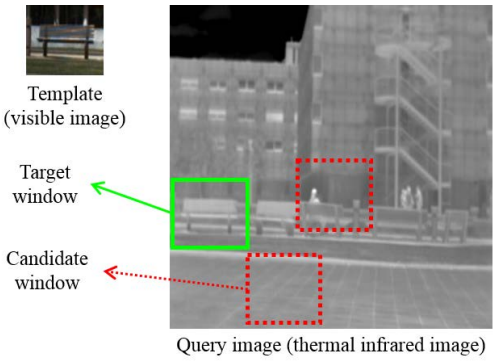


FIGURE 3. Example of template matching between visible and thermal infrared images. The sliding window mechanism compares the sim-ilarity of the template image with each candidate window of the same size as the template image. Ideally, the similarity between the template and the target window is the highest.

III. MATCHING BY SLICE TRANSFORM MATRIX MAPPING

In the proposed multisource image template matching scheme, we follow the traditional sliding window method and evaluate the maximum score between a template image and the candidate window (of the size of the template) in the query image under all possible structure features of the same gray levels. In the following, we give a general definition of the Matching by Slice Transform Matrix Mapping (MSTMM).

Assume an $m \times m$ template image is to be sought in an $n \times n$ query image as illustrated in Fig. 3. According to the traditional sliding window method, the number of candidate windows is $(n-m+1)\times(n-m+1)$. Let t be an $m\times m$ template image and w be an $m \times m$ candidate window to be compared against. Denote a translation function by $E(*)$. Thus, the translation function of $E(t)$ represents the translation that transforms the template image to the mapped template image, and the translation function of $E(w_i)$ represents the translation that transforms the candidate window image to the mapped candidate window image. Ultimately, we can find the best-matched candidate window w_* which achieves the maximum score when evaluating all candidate windows, refer to (1).

$$w_* = \underset{w_i \in U_w}{arg\,max}\{MSTMM(E(t), ES(w_i))\} \tag{1}$$

where, the set of U_w is total candidate windows in the query image, $MSTMM(*, *)$ is the matching degree measurement between the two mapped images. The measurement of

MSTMM reflects the similarity between the mapped images of template image t and the candidate window w. In order to obtain the mapped image, we first need to expand and improve the original Slice Transform (SLT) matrix.

A. THE EXPANDED SLICE TRANSFORM WITH ADAPTIVE GRAY LEVEL (EST-AGL)

The Slice Transform (SLT) was first introduced in [48] and is used as a theoretical basis for multisource image template matching in the MTM algorithm. In SLT, consider flattening an image represented as a column vector $ps = [p_1, p_2, p_3, \cdots, p_m]$ with values in the half-open interval $[a, b)$, where the pixel value of p_j denotes the gray value located at point j in the flattened image, the values of a and b are the minimum and maximum values in the flattened image, and the interval is divided into k bins. Finally, collecting the slices ps^i in columns, the SLT matrix of $S(ps)$ is defined in (2).

$$S(ps) = [ps^1, ps^2, ps^3, \cdots, ps^k] \tag{2}$$

where the slice column vector $ps^i = [p_1^i, p_2^i, p_3^i, \cdots, p_m^i]$ is an indicator function representing the entries of ps associated with the i-th bin. The value of p_j^i is defined in (3).

$$p_j^i = \begin{cases} 1 & \text{if } p_j \text{ in } i\text{-th bin} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The interval is divided into k bins, indicating that the gray level of the image is k. In [48], the gray level of the image is a fixed value defined in advance, and the gap between each gray level is equal. In order to make the transformed matrix better reflect the structural features of the image, this paper proposes to implement the adaptive gray level according to the contrast between every two pixels in the image. In the adaptive gray level principle, if the distance between the gray values of every two pixels is in the interval $(-d, d)$, the two pixels are considered to be at the same gray level. Where, the value of d is the distance threshold between the gray values of different pixels. Due to the huge difference between visible and thermal infrared images, different images can be customized with different distance thresholds to better extract the structural features of the images.

In SLT, the scale of the SLT matrix is equal to the number of gray levels, and in the Expanded Slice Transform with Adaptive Gray Level (EST-AGL), the scale of the EST-AGL matrix is equal to the number of pixels in the flattened image. We define the EST-AGL matrix $E(ps)$ as in (4).

$$E(ps) = [ps^1, ps^2, ps^3, \cdots, ps^m] \tag{4}$$

After the expanded slice transform with an adaptive gray level, the flattened image is transformed into matrix $E(ps)$. Assuming that an image with 256 gray levels is converted into an image with 15 gray levels, according to the slice transform in [48], the 15 gray levels are defined by $\alpha = [0, 17, 34, 51, 68, \cdots, 187, 204, 221, 238, 256]$. Each element in α is the boundary value between gray levels, and the difference between adjacent elements is seventeen. Assuming that the distance threshold between the gray values of different pixels is seventeen in the expanded slice transform with adaptive gray level. According to the above assumptions, the difference between matrix $S(ps)$ and $E(ps)$ is shown in Fig. 4.

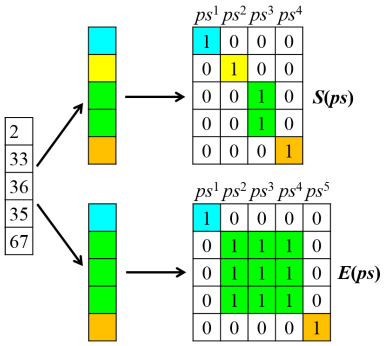


FIGURE 4. The matrix $S(ps)$ of SLT and the matrix $E(ps)$ of EST-AGL for a 5-pixel flattened image patch have five gray values. The height of the SLT matrix is equal to the number of pixels, the width of the SLT matrix is equal to the number of gray levels; the height and width of the EST-AGL matrix are both equal to the number of pixels. Different background colors indicate different gray levels at corresponding positions.

B. THE GRAY LEVEL ANALYSIS AND PATCH SEGMENTING MECHANISM

The matrix of EST-AGL reflects the spatial distribution of the same gray levels. We will analyze the spatial structure of the image in combination with the EST-AGL matrix in this part.

For the EST-AGL matrix $E(ps)$ for a 2-pixel flattened image patch, we know that it has at most two gray levels and its corresponding EST-AGL matrices have only two types, as illustrated in Fig. 5a. Similarly, the 3-pixel flattened image patch has at most three gray levels and five types of EST-AGL matrices, and the 4-pixel flattened image patch has at most four gray levels and fifteen types of EST-AGL matrices, as illustrated in Fig. 5b and Fig. 5c.

A complete flattened image contains too many different types of EST-AGL matrices, so it must be segmented. Segmenting a complete image into patches of the same size can greatly reduce the execution time of the algorithm, thereby improving the efficiency of the algorithm.

Image segmentation can be accomplished through two mechanisms. The first segmentation mechanism, which is also the segmentation mechanism chosen by most algorithms, is isometric segmentation. For example, an $a \times b$-sized image can be segmented into $a/c \times b/d$ patches of $c \times d$-sized, as shown in the subfigures of the four corners in Fig. 6. The second segmentation mechanism is the sliding window segmentation. For example, an $a \times b$-sized image can be segmented into $(a-c+1) \times (b-d+1)$ patches of $c \times d$-sized, as shown in all nine subfigures in Fig. 6. Obviously, the fewer the patches are, the faster the program executes,

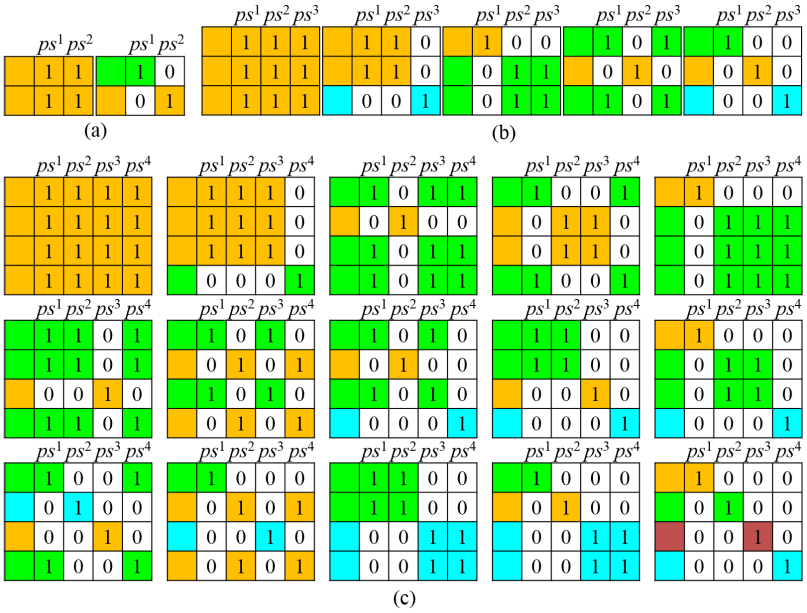


FIGURE 5. The example of EST-AGL matrix for all possible gray level cases: (a) 2-pixel flattened image patch; (b) 3-pixel flattened image patch; (c) 4-pixel flattened image patch. Each matrix diagram of EST-AGL is a simplification of the EST-AGL matrix diagram in Fig. 4. The left column of each matrix diagram of EST-AGL is the flattened image with different gray levels, followed by the EST-AGL matrix corresponding. Different background colors indicate different gray levels at corresponding positions.

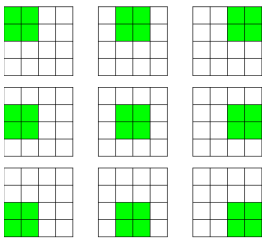


FIGURE 6. The example of the segmentation mechanism. In the sliding window segmentation mechanism, a 4 × 4-sized image is segmented into 3 × 3 patches of 2 × 2-sized. In the isometric segmentation mechanism, a 4 × 4-sized image is segmented into 2 × 2 patches of 2 × 2-sized, corresponding to the top left, top right, bottom left, and bottom right patches in this figure.

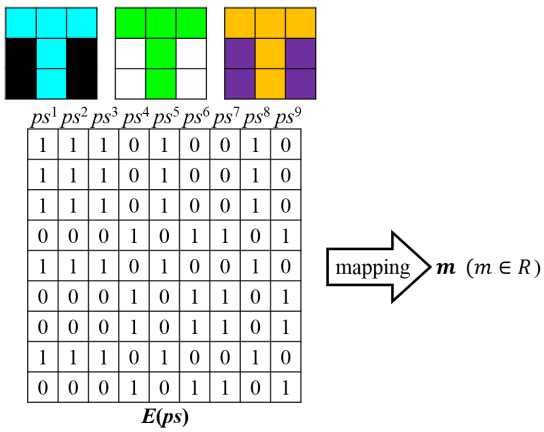


FIGURE 7. The three different "T" structured image patches with different gray levels and corresponding same matrix $E(ps)$. At the top of each diagram is three different "T" structured image patches, followed by the same EST-AGL matrix corresponding, and the rightmost m is the real value to which this EST-AGL matrix is mapped.

conversely, the more the patches are, the finer the spatial structure involved in the calculation is, and the more robust the algorithm is. To balance performance and efficiency, the scheme in this paper prefers the sliding window segmentation mechanism.

C. MSTMM SIMILARITY MEASURE BY EST-AGL MATRIX MAPPED IMAGE

The theory of EST-AGL transforms flattened image patches with different gray levels into a matrix $E(ps)$, and we can measure the similarity of two image patches by comparing the difference between the gray levels mapped by the matrix $E(ps)$. As shown in Fig. 7, the $E(ps)$ matrix associated with these "T" structured image patches with different gray levels of three different modalities are the same. Since the same EST-AGL matrix can be mapped to exactly the same real numbers, and the difference between the same real numbers is 0, these "T" structured image patches of different modalities are exactly the same. After all the patches in the image are mapped to different real values through the EST-AGL matrix, we can get the mapped image of this image. As shown in Fig. 8, a 3×3-sized "T" structured image is finally converted into a 2×2-sized mapped image. Finally, the purpose of measuring the similarity between the heterologous images

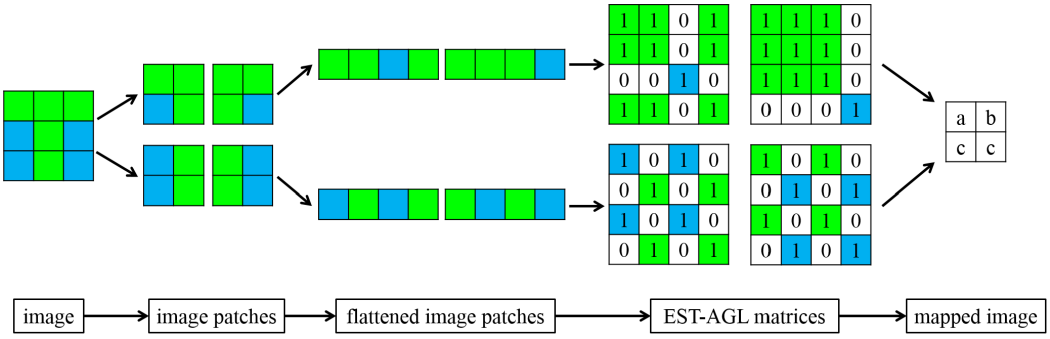


FIGURE 8. The image mapping process: first, a 3 × 3-sized "T" structured image is segmented into 2 × 2 patches of 2 × 2-sized by a sliding window segmentation mechanism; then, these patches are flattened and their corresponding EST-AGL matrices are obtained; finally, a 2 × 2-sized mapped image is obtained according to the principle of mapping the same EST-AGL matrix to the same real number.

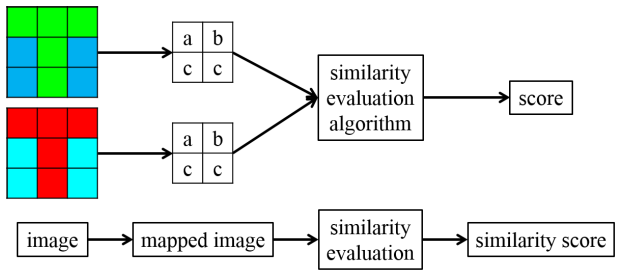


FIGURE 9. The process of MSTMM similarity measuring: first, the input images are converted into their corresponding mapped images through EST-AGL matrix transfer; then, the homologous image similarity evaluation algorithm is used to evaluate the similarity between the mapped images from heterologous images; finally, the similarity score between the mapped images is used as the similarity score between the heterologous images.

can be achieved by using the homologous image similarity evaluation algorithm on the mapped images of these heterologous images. The complete process of the MSTMM similarity measurement is shown in Fig. 9. In Fig. 9, a pair of "T" structured heterologous images are converted into a real mapped image with the same modality, and since the mapped images are completely the same, this pair of heterologous images are exactly the same.

From the similarity measuring process of MSTMM shown in Fig. 9, we can conclude that the EST-AGL matrix mapping mechanism and the similarity evaluation algorithm will directly determine the performance of the MSTMM scheme.

The mapping mechanism of EST-AGL matrix is the most important in the MSTMM scheme. The quality of the mapped image will directly determine the evaluation performance of the similarity evaluation algorithm. The simplest mapping mechanism is to directly map the EST-AGL matrix to a random integer, which we call Matching by Slice Transform Matrix Mapping of Integer Mapping (MSTMM-IM), refer to (5), where m is a random integer. The mechanism of MSTMM-IM has a high mapping efficiency because of its very simple mapping mechanism. However, the mapped integer of the EST-AGL matrix is only a representation of this matrix, and different integers only represent different EST-AGL matrices, and the difference between integer values is meaningless and cannot represents the degree of similarity between different EST-AGL matrices. To solve this problem, this paper develops a neural network mapping mechanism we call Matching by Slice Transform Matrix Mapping of Network Mapping (MSTMM-NM), refer to (6), where w is the weight to be trained. In the following, we will demonstrate how to design MSTMM-IM and MSTMM-NM based on specific examples.

$$MSTMM\text{-}IM(E(ps)) = \{m \mid m \in \mathbb{Z}^+\} \tag{5}$$

$$MSTMM\text{-}NM(E(ps)) = \{w \times m \mid m \in \mathbb{Z}^+\} \tag{6}$$

1) MSTMM-IM MECHANISM

The mapping mechanism of MSTMM-IM is the simplest EST-AGL matrix mapping mechanism. Each type of EST-AGL matrix can theoretically be mapped to an arbitrary integer, but considering these issues, such as the convenience of mapping expression, as simple an algorithm as possible to implement the matching scheme in the occlusion scene, the difference between mapped values, the gray level distribution of image patches under most occlusions, the similarity between EST-AGL matrices, etc., the simplified mapping process of the MSTMM-IM mechanism combined with Fig. 5 and Fig. 8 is shown in Fig. 10. The simplified mapping process and the mapped results of the MSTMM-IM mechanism for a 3-pixel flattened image patch with different gray levels and a 4-pixel flattened image patch with different gray levels are shown in Fig. 10, respectively. Since the image patches in most occlusion scenes have fewer gray levels, the matrix of EST-AGL corresponding to the flattened image patch with fewer gray levels is mapped to an integer value as close to the median as possible. In particular, the matrix of EST-AGL corresponding to the flattened image patch with only one gray level is mapped to the median value, as shown in the position of the red dashed box in Fig. 10. This approach can reduce the difference between the median value and the other integer values, thereby increasing the similarity

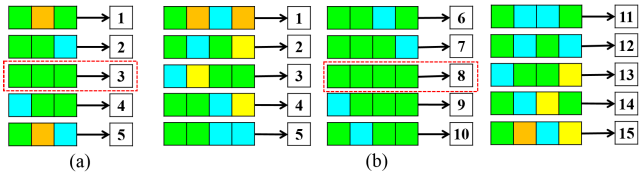


FIGURE 10. The example of MSTMM-IM mechanism for all possible gray level cases: (a) 3-pixel flattened image patch; (b) 4-pixel flattened image patch. The left side of each pair of mappings in the figure is the flattened image patches with different gray levels, and the right side is the corresponding mapped values. The mapping process from the left to the right can be processed according to Fig. 8. The matrix of EST-AGL corresponding to each flattened image patch can be derived from Fig. 5.

between the image patches in the normal scene and the image patches in the occlusion scene, thereby increasing the similarity between the candidate window in the normal scene and the candidate window in the occlusion scene. The similarity score of the MSTMM-IM mapping mechanism can reduce the impact of occlusion on the similarity evaluation algorithm, so that we can have more choices when choosing the similarity evaluation algorithm, and finally have a better balance in the performance and efficiency of the MSTMM-IM mechanism.

2) MSTMM-NM MECHANISM

Compared to image patches with more pixels, there are fewer gray levels and fewer types of corresponding EST-AGL matrices in 3-pixel image patches and 4-pixel image patches, and it is possible to manually assign a unique mapping value to each particular type of EST-AGL matrix. Therefore, for multi-pixel image patches with more gray levels and more types of EST-AGL matrix, we introduce a neural network to automatically assign mapping values to every EST-AGL matrix. For the convenience of program processing, we can first randomly assign a different integer to different EST-AGL matrices as mapping values, and these mapping values can be used as the input of the MSTMM-NM mechanism according to (6). The number of neurons in the MSTMM-NM mechanism is equal to the number of random assigned distinct integer values. The output of each neuron in MSTMM-NM is the mapped value of the MSTMM-NM mechanism for the EST-AGL matrix corresponding to the integer value of the input of the MSTMM-NM network. The mapped image of the MSTMM-NM mechanism can be built by replacing the corresponding integer value in the random mapped image of the input visible and thermal infrared images with the output values of the neurons in the MSTMM-NM network. It follows that the MSTMM-NM network is designed as a fully connected feedforward neural network (FNN) is sufficient.

The similarity between the heterogeneous images can be evaluated by feeding the input visible and thermal infrared images into the similarity evaluation algorithm after MSTMM-NM processing. The MSTMM-NM network can be trained by minimizing the distance of the input paired visible and thermal infrared image pairs and maximizing the distance of the input unpaired visible and thermal infrared image pairs. Since the existence of the fewer texture features of thermal infrared images than visible images, in continuous learning by minimizing the distance of the input paired visible and thermal infrared image pairs, the similarity between normal scene image patches and occluded scene image patches can be increased so that the impact of occlusion on the subsequent similarity evaluation algorithm is reduced.

For the similarity evaluation algorithm, since the heterologous images have been converted into grayscale-independent homologous images, each homologous image similarity evaluation algorithm can theoretically be used. Considering the problem that the functions involved in the backpropagation process required by the neural network must be continuous, this paper prefers the SSD algorithm to achieve similarity evaluation during network training. Compared with other similarity evaluation algorithms, although the performance of the SSD algorithm is weak, the SSD algorithm is simple and high efficient. In addition, since the SSD algorithm is more sensitive to the individual outliers in the matching pair, the distance between the outliers in the matching pair can be reduced by continuous learning and adjustment, thereby increasing the similarity score of the matching pair. For the specific deployment method of the SSD algorithm in the network, please refer to the processing in the next subsubsection.

In summary, combined with the mapping principle of Fig. 10, we can introduce the network structure under the 2-pixel image patch and the 3-pixel image patch with different gray levels, as shown in Fig. 11a and 11b, respectively. The mapped image 1 and image 2 in Fig. 11 are the random mapped image of the input visible and thermal infrared images, and the output image by the mixer is the mapped image of the MSTMM-NM mechanism. The function of Mixer is to replace the integer values in the input mapped image with the real values processed by the neurons.

The design of the loss function is a crucial aspect of neural network design, as it is directly connected to network training and inference performance. The loss function in a fully connected FNN is generally defined by the distance between the network output value and the label value; however, the matching degree score of the template matching task output does not have a fixed value, i.e., there is no label value. Therefore, in this paper, the multi-classification cross-entropy loss function from the CNN classification task is introduced into this fully connected FNN, refer to (7).

$$L = -\sum_{i=1}^{K} y_i \, log(p_i) \tag{7}$$

where, the value of K is the number of categories. The variable of y_i is the label value of the category, that is, if the category is i, then y_i is equal to 1, otherwise it is equal to 0. The variable of p_i is the final output of the neural network, that is, the probability of that category is p_i. The variable of p_i is derived by computing the output of the neural network using the softmax function.

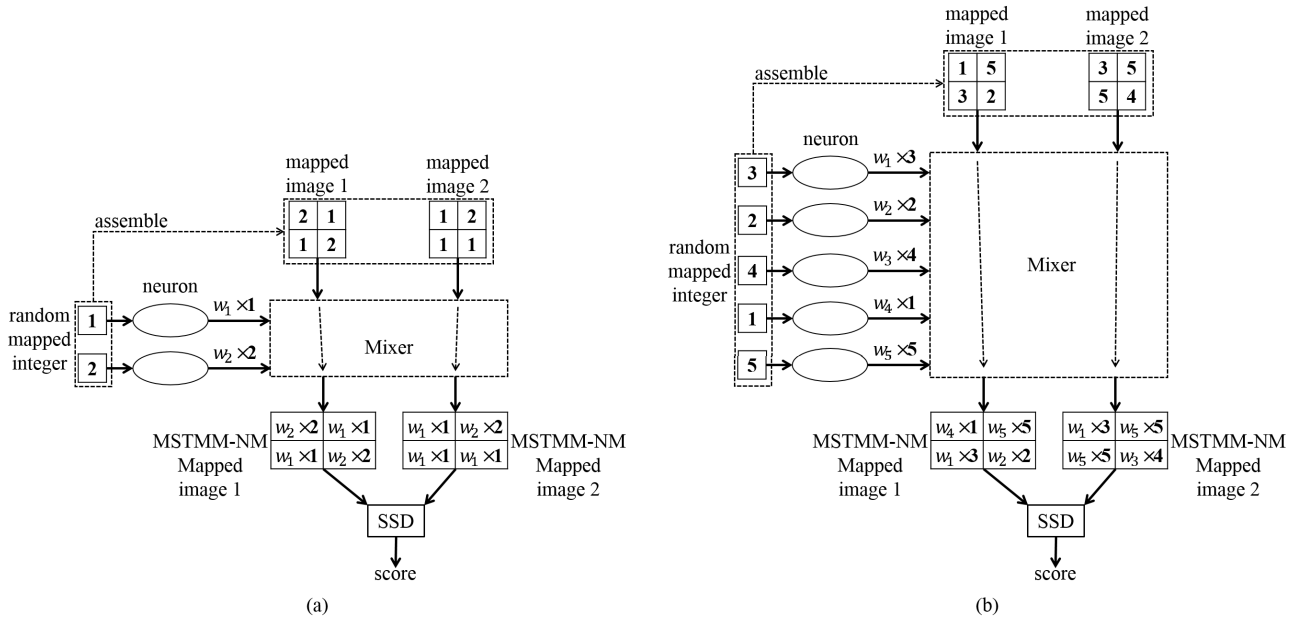


FIGURE 11. The example of MSTMM-NM mechanism architecture for all possible mapped integer values: (a) 2-pixel flattened image patch; (b) 3-pixel flattened image patch. The input of the network includes the mapped image corresponding with the heterologous image assembled from the random mapped integer and the integer value corresponding with the EST-AGL matrix random mapped. The function of Mixer is to replace the integer values in the input mapped image with the real values processed by the neurons. The mapped image processed by the Mixer is the MSTMM-NM mapped image. The output score of the architecture is the similarity score between different MSTMM-NM mapped images.

For the template matching task, the most important thing to train the network using the classification loss function is how to classify the dataset and decide how many categories there are. This paper is the first innovation to use subwindows at different locations in an image to classify between subwindows, thus solving the long-standing problem of how to use neural network learning to solve the template matching task. Since the evaluation criterion of the template matching task is the accuracy of the template matching, which is closely related to the matching position, and the current neural network learning cannot solve the position-related problems in the template matching task, the only way to realize the use of neural network learning to solve the template matching problem is to transform the matching position problem into a classification problem.

Why is it possible to convert the template matching location problem into a classification problem? The essence of the classification problem is that a certain category of data is processed by the network to generate a score value, and the score value is processed into the probability value of the data belonging to each category, where the category with the largest probability value is the category of the data, so as to realize the classification of the data. Since the output of the neural network designed in this paper also has a similar score value, the same method of classification neural network can be used to process this score value.

How to divide the categories of templates? A template is taken at a certain interval size on the base image, and the template at each position is used as a classification category, thus achieving the division of template categories. The interval size should not be too small, because too small interval size will lead to too many templates and thus too many classification categories, which is not conducive to the convergence of network training. And the interval size should not be too large, because too large an interval size will make the difference between templates too big so that the trained network will reduce the accuracy of template matching.

3) SIMILARITY EVALUATION

In this paper, the similarity evaluation algorithm is required in both the MSTMM-NM network training process and the MSTMM template matching scheme. During the training process of the MSTMM-NM network, the similarity evaluation algorithm should be able to reduce the distance between outliers in matching pairs through continuous learning. From previous subsubsection, this paper prefers the SSD algorithm to achieve similarity evaluation during the training process of the MSTMM-NM network. For the similarity evaluation algorithm used in the MSTMM template matching scheme, since the heterologous images have been converted into grayscale-independent homologous images, each homologous image similarity evaluation algorithm can theoretically be used, but to balance the performance and efficiency of the algorithm, this paper prefers the NCC algorithm to achieve similarity evaluation.

Refer to (1), the similarity evaluation can be calculated according to (8) and (9) in the MSTMM-NM network training process and the MSTMM template matching scheme.

The signs of $m(t)$ and $m(w_i)$ in (8) and (9) stand for the mapped images of template and subwindow.

$$MSTMM\text{-}NM(E(t), ES(w_i)) = SSD(m(t), m(w_i)) \quad (8)$$

$$MSTMM(E(t), ES(w_i)) = NCC(m(t), m(w_i)) \quad (9)$$

For the SSD similarity evaluation algorithm used in the training process, this paper uses the convolution operation to improve the efficiency of the SSD algorithm. The SSD formula is expanded in (10).

$$\begin{aligned} SSD(m(t), m(w_i)) &= \sum_{i=1}^{p}\sum_{j=1}^{q}(m(t) - m(w_i))^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q}(m(t)^2 + m(w_i)^2 - 2m(t)m(w_i)) \\ &= \sum_{i=1}^{p}\sum_{j=1}^{q}m(t)^2 + \sum_{i=1}^{p}\sum_{j=1}^{q}m(w_i)^2 \\ &\quad - 2\sum_{i=1}^{p}\sum_{j=1}^{q}m(t)m(w_i) \\ &= sum(m(t)^2) + sum(m(w_i)^2) \\ &\quad - 2sum(m(t)m(w_i)) \end{aligned} \quad (10)$$

The values of p and q in (10) represent the length and width of the mapped image $m(t)$, respectively, and the sign of $m(w_i)$ stands for the mapped image of the candidate image window in the query image. Where the function of $sum()$ represents the sum operation of the matrix of the mapped image. In summary, refer to (1), the summation operations related to the candidate window w_i in (10) can be implemented according to the convolution operations of (11) and (12).

$$\{sum(m(w_i)^2) \mid w_i \in U_w\} = \{ones(p, q) \otimes Q^2\} \quad (11)$$

$$\{sum(m(t)m(w_i)) \mid w_i \in U_w\} = \{m(t) \otimes Q\} \quad (12)$$

For the NCC similarity evaluation algorithm used in the MSTMM template matching scheme, for convenience, this paper directly uses the function of *matchTemplate*() in the OpenCV library.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we will conduct a detailed experimental study to test the performance of the MSTMM scheme. We first test the performance of the MSTMM-IM mechanism under different configurations, second test the performance of the MSTMM-NM mechanism under different configurations, third compare the performance of the MSTMM scheme with other template matching algorithms, and finally test the execution efficiency and training efficiency of the MSTMM scheme with other template matching algorithms.

The algorithms involved in the comparison experiments include the MSTMM-IM mechanism and the MSTMM-NM mechanism in the MSTMM scheme, the most common template matching algorithm NCC, the algorithm of MTM which is the source of the algorithm ideas in this paper, the traditional visible and thermal infrared image template matching algorithm LCTS, the weakly supervised deep metric learning network NCCNet based on CNN template matching, the multimodal image GAN model MUNIT, the template matching algorithm of M-RGBIR for visible and thermal infrared image based on densely connected CNN, and the latest visible and thermal infrared image template matching algorithm STVO.

The dataset used in our experiments is the same as that used in [50], which we call the Log-Gabor Histogram Descriptor (LGHD) dataset. The LGHD dataset is a rarely registered dataset that contains 44 pairs of visible and thermal infrared images. All images in the dataset are pre-scaled to 256×256. All the query images used in our experiment are the 256×256 thermal infrared images. The size of all template images is 64×64, and the template image is sliding extracted from the 256×256 visible images at certain size intervals. The number of template images that can be extracted from a base image can be calculated according to (13).

$$n_{tem} = ((qH - tH)//\mu + 1) \times ((qW - tW)//\mu + 1) \quad (13)$$

where the value of n_{tem} is the number of template images extracted, the sign of "//" indicates rounding down, the value of μ is the interval size, the values of qH and qW are the height and width of the base image, and the values of tH and tW are the height and width of the template image. For example, for a 256×256 visible base image, the size of the template image is 64×64, the interval size is set to 8, and eventually, a total of 625 template images can be extracted.

In order to test the performance of the MSTMM scheme in real scenes with different occlusion degrees, this paper first simulates the different occlusion scenarios by generating rectangular spots of random size, fixed rotation angle, random position, and random grayscale on the thermal infrared query image in the test image pair. Then, the target region of the unoccluded visible template image is searched on the thermal infrared query image with different occluded degree scenes. In order to simulate the real occlusion scene, the pixel values of the pixels inside the rectangular spot are not pure, but slightly different, as shown in Fig. 12. These random slightly different image spots are achieved by saving rectangular pure value spots as ".jpg" files. The variation range of the pixel value of the pixels inside the rectangular spot is random, which is determined by the codec scheme of the ".jpg" file inside the OpenCV library. In this paper, according to the size of the randomly generated rectangular spots, the degree of occlusion is divided into four levels, namely, level 0 without occlusion, level 1 with less occlusion, level 2 with greater occlusion, and level 3 with maximum occlusion. Where the random length of the rectangular spot for level 1 ranges from 6 to 12 pixels and the random width ranges from 3 to 9 pixels, the random length for level 2 ranges from 9 to 15 pixels and the random width ranges from 6 to 12 pixels, the random length for level 3 ranges from 12 to 20 pixels and the random width ranges

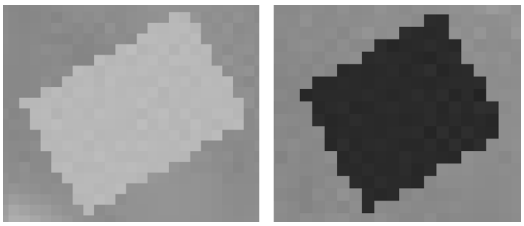


FIGURE 12. Detail of the interior of the rectangular spot. The pixel values of the pixels inside the rectangular spot are not pure, but slightly different.

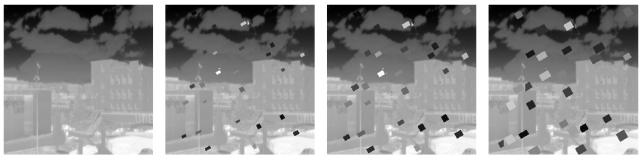


FIGURE 13. The example of thermal infrared images and their occlusion scenes with different occlusion levels.

from 9 to 16 pixels. The specific size of the rectangular spot is shown in Fig. 13.

In our experiments, we evaluate an algorithm using two criteria: matching Success Rate and algorithm runtime. We first define the matching accuracy, which is defined in (14).

$$accuracy = \frac{area(B_{gt} \cap B_{pred})}{area(B_{gt})} \tag{14}$$

where, the signs of B_{pred} and B_{gt} denote the corresponding predicting bounding box and ground truth bounding box, respectively. When the matching accuracy exceeds the threshold, we consider the template matching to be successful. We set the threshold value at 0.6. The **Success Rate** is defined as the number of successfully matched templates divided by the total number of templates matched.

The device configuration used in these experiments is an Intel(R) Core(TM) i3-3110M CPU of 2.4 GHz, a memory of 4 GB, the Windows 7 operating system, and the Python 3.7 development environment. The third-party libraries used in this paper are Numpy library version 1.21.1 and OpenCV library version 4.5.3.56, and the neural network training libraries are Pytorch library version 1.8.0 and Tensorflow library version 1.15.5. Note that the NCC algorithm used in the MSTMM scheme is implemented by the *matchTemplate*() function in the OpenCV library.

A. MSTMM-IM MECHANISM PERFORMANCE ANALYSIS

This performance analysis experiment of the MSTMM-IM mechanism is mainly divided into three experiments. Firstly, the performance of the MSTMM-IM mechanism under different distance thresholds between the gray values of different pixels is tested. Second, test the performance of the MSTMM-IM mechanism under different occlusion levels. The third is to test the performance of the MSTMM-IM mechanism under different distance thresholds and occlusion levels. The range of the distance threshold of the visible template image and the thermal infrared query image in Experiment 1 and Experiment 3 is $d = [1, 2, 3, 4, 5, 6, 7, 8, 9]$. The test set is the 44 pairs of the visible and thermal infrared images in the LGHD dataset. The four copies of thermal infrared images in the test set are constructed to build four test sets with four levels of occlusion. In the test set, the interval size is set to 32, and a total of 49 template images can be extracted from one visible image, i.e., the test set contains 2156 pairs of template and query image pairs.

1) EXPERIMENT 1: DIFFERENT DISTANCE THRESHOLDS IN MSTMM-IM

This experiment will evaluate the performance of the MSTMM-IM mechanism for different image patches at different distance thresholds d between the gray values of different pixels in the first subsection of Section III. The image patches are 1×3, 3×1, 1×4, 4×1, and 2×2, respectively. As in Fig. 14, show the statistical results of the success rate when the degree of occlusion is the greater occlusion at level 2.

The general rule is shown in Fig. 14, the MSTMM-IM mechanism can achieve better matching results when the degree of occlusion is the greater occlusion at level 2. Since both patch 1×3 and patch 3×1 contain three pixels and are all pixels taken in the single-dimensional direction of the original image, the trends of these polylines changes are almost the same, and the results for patch 1×4 and patch 4×1 are similar. Although patch 2×2 also contains four pixels, it is a patch whose pixels are taken in two dimensions of the original image, so the trend of these polylines changes is different from those of patch 1×4 and patch 4×1. As shown in Fig. 14, patch 1×3 and patch 3×1 match best when the distance threshold of the visible template image is four and the distance threshold of the thermal infrared query image is two; patch 1×4 and patch 4×1 match best when the distance threshold of the visible template image is two and the distance threshold of the thermal infrared query image is one; patch 2×2 matches best when the distance threshold of the visible template image is nine and the distance threshold of the thermal infrared query image is five. These results can also prove that the visible image has rich texture and the thermal infrared image has less texture, therefore, the larger distance threshold of the visible image can reduce the texture to a certain extent and make it match better with the thermal infrared image which has less texture. Next, we will test the performance of the MSMM-IM mechanism for different image patches at different occlusion levels using the distance thresholds of the visible template image and the thermal infrared query image when the image patches with the best matching result achieve the best matching performance.

2) EXPERIMENT 2: DIFFERENT OCCLUSION LEVELS IN MSTMM-IM

In this experiment, the distance thresholds of patch 1×3 and patch 3×1 were four for visible template images and two

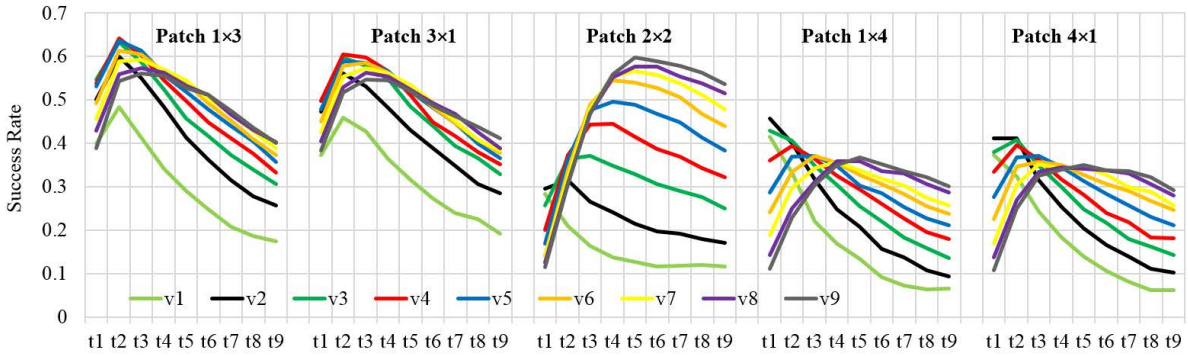


FIGURE 14. The success rate for different patches (see title) and different distance thresholds (see legend and horizontal axis). In the figure, the signs of "*v* + *num*" and "*t* + *num*" represent the distance threshold of the image, where the sign of *v* denotes the visible image, the sign of *t* denotes the thermal infrared image, and the sign of *num* is the distance threshold.

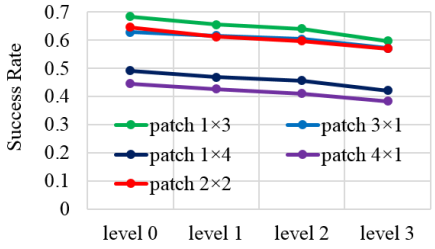


FIGURE 15. The success rate for different patches (see legend) and different occlusion levels (horizontal axis).

for thermal infrared query images; the distance thresholds of patch 1 × 4 and patch 4 × 1 were two for visible template images and one for thermal infrared query images; the distance threshold of patch 2 × 2 is nine for visible template images and five for thermal infrared query images. As in Fig. 15, show the statistical results of the success rate for four different levels of occlusion degree.

The general rule is shown in Fig. 15, the performance of the MSTMM-IM mechanism varies within 10% at most in different occlusion degrees, which is within an acceptable range. The patch 1 × 3 achieves the best matching result when the original image pixels are taken in a single-dimensional direction. The patch 2 × 2 achieves the best matching results when the original image pixels are taken in two dimensions of the original image. Next, we will use patch 1 × 3 with the best matching results to explore the trends in the matching performance of the MSTMM-IM mechanism at different distance thresholds and occlusion levels.

3) EXPERIMENT 3: DIFFERENT DISTANCE THRESHOLDS AND OCCLUSION LEVELS IN MSTMM-IM

As in Fig. 16, show the statistical results of the success rate in different distance thresholds and occlusion levels.

The general rule is shown in Fig. 16, patch 1 × 3 had the same trend of polylines changes in different occlusion levels in the MSTMM-IM mechanism. The results suggest that different occlusion degrees only affect the matching performance of the MSTMM-IM mechanism, but cannot change the trend of matching performance.

B. MSTMM-NM MECHANISM PERFORMANCE ANALYSIS

This performance analysis experiment of the MSTMM-NM mechanism is mainly divided into four experiments: the first is to test the performance of the MSTMM-NM mechanism under different distance thresholds; the second is to test the performance of the MSTMM-NM mechanism under different interval sizes; the third is to test the effect of the size of the training set on the performance of the MSTMM-NM mechanism; the fourth is to test the performance of the MSTMM-NM mechanism under different occlusion levels.

In the first two experiments, the visible and thermal infrared image pair used in the training set is shown in Fig. 17a, the test set is the remaining 43 pairs of the visible and thermal infrared images in the LGHD dataset. In the last two experiments, the training set contains one, two, three, and four pairs of visible and thermal infrared images, respectively, which corresponds to Fig. 17a, Fig. 17ab, Fig. 17abc, and Fig. 17abcd. The test set is the remaining 40 pairs of the visible and thermal infrared images in the LGHD dataset in the last two experiments. All training sets are datasets of unoccluded scenes. The test sets in the first three experiments are datasets with the greater occlusion at level 2. In the fourth experiment, the test set is the dataset under four scenarios with different occlusion degrees. In the test set, the interval size is set to 32, and a total of 49 template images can be extracted from one visible image, that is, the test set of the first two experiments contains 2107 pairs of template and query image pair, and the last two experimental test set contains 1960 pairs of template and query image pair.

1) EXPERIMENT 4: DIFFERENT DISTANCE THRESHOLDS IN MSTMM-NM

In this experiment, the interval size of the visible image in the training set is set to eight, and a total of 625 template images

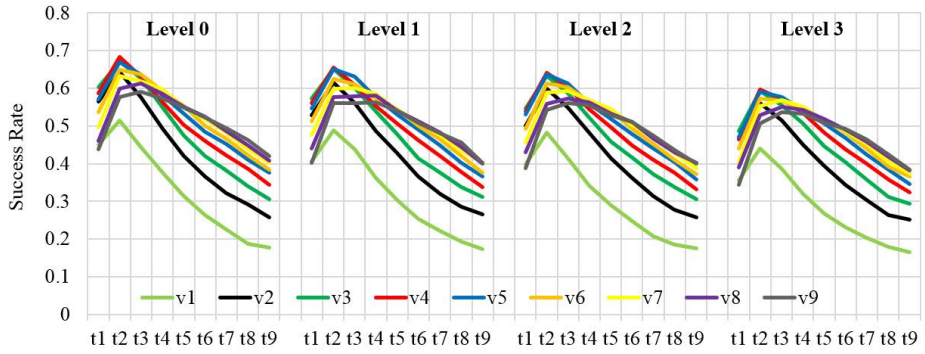


FIGURE 16. The success rate for different occlusion levels (see title) and different distance thresholds (see legend and horizontal axis). In the figure, the signs of "*v* + *num*" and "*t* + *num*" represent the distance threshold of the image, where the sign of *v* denotes the visible image, the sign of *t* denotes the thermal infrared image, and the sign of *num* is the distance threshold.

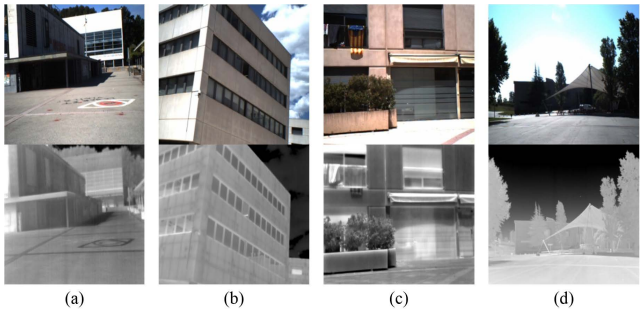


FIGURE 17. The visible and thermal infrared image pairs used in the training of MSTMM-NM mechanism performance analysis experiments. Image pair (a) is used in experiments 4 to 7, and image pairs (b)–(d) are used in EXPERIMENT 6 and EXPERIMENT 7. The top image of each pair is a visible image, and the bottom image is a corresponding thermal infrared image.

can be extracted, i.e., the training set contains 625 pairs of template and query image pairs.

This experiment will evaluate the performance of the MSTMM-NM mechanism for different image patches at different distance thresholds d between the gray values of different pixels in the first subsection of Section III. The range of the distance threshold of the visible template image and the thermal infrared query image is $d = [1, 2, 3, 4, 5, 6, 7, 8, 9]$. The image patches are 1×3, 2×2, 2×3, and 3×2, respectively. As in Fig. 18, show the statistical results of the success rate when the degree of occlusion is the greater occlusion at level 2.

Comparing the patch 1×3 and patch 2×2 results in Fig. 14, the results shown in Fig. 18 suggest that the MSTMM-NM mechanism with a neural network structure outperforms the MSTMM-IM mechanism in most cases, thus supporting our viewpoint in the third subsection of Section III. Since both patch 2×3 and patch 3×2 contain six pixels and are all pixels taken in two dimensions of the original image, the trends of these polylines changes are almost the same under different distance thresholds. The conclusion that the polylines changes of the MSTMM-NM mechanism in the same way of taking pixels and the same number of pixels are roughly the same indicates that the MSTMM-NM mechanism has similar properties compared to the MSTMM-IM mechanism. The light green polyline in Fig. 18 indicates that, regardless of the image patches structure, the MSTMM-NM mechanism does not achieve better results when the distance threshold of the visible image with rich texture is smaller than the distance threshold of the thermal infrared image with less texture. The MSTMM-IM mechanism also has similar properties in terms of distance thresholds for visible images and thermal infrared images. The results of comparing other image patches with patch 1×3 shown in Fig. 18 suggest that image patches with more pixels and more complex structures may not necessarily achieve better matching results. Next, we will test the performance of the MSMM-NM mechanism for different image patches at different interval sizes using the distance thresholds of the visible template image and the thermal infrared query image when the image patches with the best matching result achieve the best matching performance.

2) EXPERIMENT 5: DIFFERENT INTERVAL SIZES IN VISIBLE IMAGE OF TRAINING SET

In this experiment, the interval size in the visible image used in the training set is set between 4 and 64 every six pixels; the distance threshold of patch 1×3 is five for visible template images and five for thermal infrared query images; the distance threshold of patch 2×2 is nine for visible template images and three for thermal infrared query images; the distance threshold of patch 2×3 is nine for visible template images and two for thermal infrared query images; the distance threshold of patch 3×2 is six for visible template images and three for thermal infrared query images. As in Fig. 19, show the statistical results of the success rate for this experiment with different patch sizes and different interval sizes of the visible image.

According to the experimental results of patch 1×3 compared with other image patches shown in Fig. 19, the matching performance of image patches with more pixels and more complex structures is not stable in resisting the

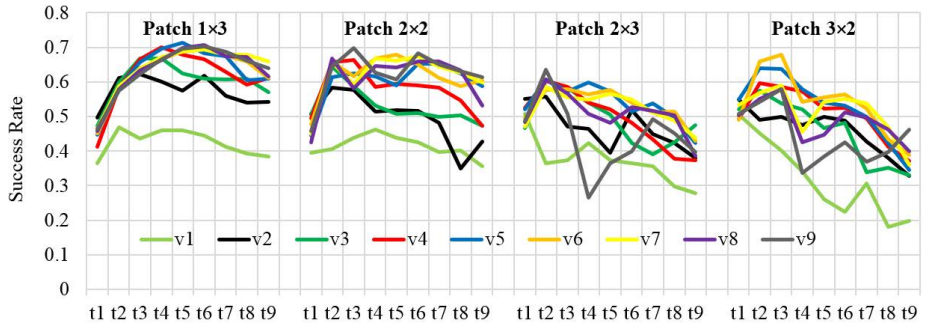


FIGURE 18. The success rate for different patches (see title) and different distance thresholds (see legend and horizontal axis). In the figure, the signs of "*v* + *num*" and "*t* + *num*" represent the distance threshold of the image, where the sign of *v* denotes the visible image, the sign of *t* denotes the thermal infrared image, and the sign of *num* is the distance threshold.

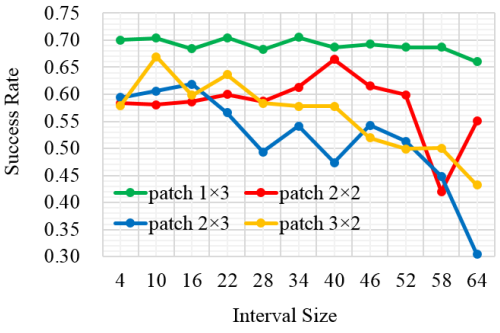


FIGURE 19. The success rate for different patches (see legend) and different interval sizes (horizontal axis) in the visible image used in the training set.

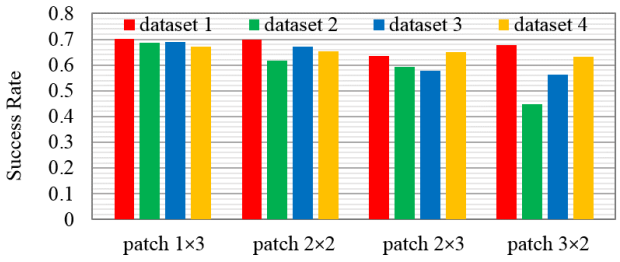


FIGURE 20. The success rate for different patches (horizontal axis) and different sizes of the training set (see legend).

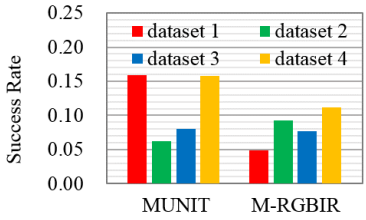


FIGURE 21. The success rate for different algorithms (horizontal axis) and different sizes of the training set (see legend).

change of interval size. In terms of matching performance, patch 1 × 3 can not only resist the change of interval size but also has higher performance than other image patches. Combining the experimental results of this experiment and Experiment 4, we will set the interval size to eight to conduct the next experiment with different sizes of the training set.

3) EXPERIMENT 6: DIFFERENT SIZES OF TRAINING SET

In this experiment, the distance threshold for different patches is the same as in Experiment 5. As in Fig. 20, show the statistical results of the success rate for this experiment with different patch sizes and different sizes of the training set.

As shown in Fig. 20, more training sets sometimes did not improve the performance of the MSTMM-NM mechanism. We speculate that the reason for this phenomenon may be that the newly added image pairs are negative samples for the original training set due to the huge difference in shooting time, weather conditions, and shooting scenes of the four pairs of visible and thermal infrared image pairs. This problem can be solved by filtering the training set before training. As shown in Fig. 21, this anomaly also occurs when M-RGBIR and MUNIT neural network algorithms are trained using the unfiltered training set (the training set used in this experiment). Note that the MUNIT statistical results are obtained under the assumption that the model converges when trained for 1000 epochs; the M-RGBIR statistical results are obtained under the assumption that the model converges when trained for 30 epochs. Next, we will explore the matching performance of the MSTMM-NM mechanism under different occlusion levels based on the results obtained in this experiment and the previous two experiments.

4) EXPERIMENT 7: DIFFERENT OCCLUSION LEVELS IN MSTMM-NM

In this experiment, the interval size of the visible images in the training set will be set to eight, the distance threshold for different patches is the same as in Experiment 5. Only the training set of patch 2 × 3 contains four pairs of visible and thermal infrared images, and the training sets of other

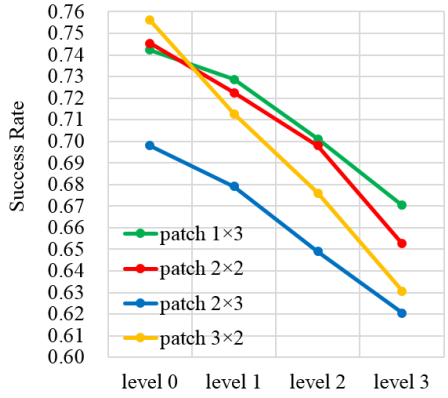


FIGURE 22. The success rate for different patches (see legend) and different occlusion levels (horizontal axis).

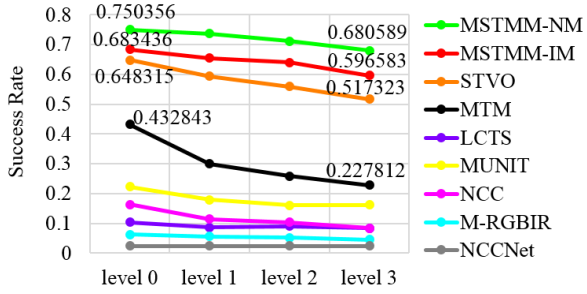


FIGURE 23. The success rate for different algorithms (see legend) and different occlusion levels (horizontal axis). The tag value is the success rate of different algorithms under occlusion levels of zero and three.

patches only contain one pair of the visible and thermal infrared images. As in Fig. 22, show the statistical results of the success rate for this experiment with different patch sizes under four different occlusion levels.

The general rule is shown in Fig. 22, the minimum performance varies of the MSTMM-NM mechanism under different occlusion levels is about 7%, which is within an acceptable range. As shown in Fig. 22, in the occlusion scene of level 0 (no occlusion), patch 3×2 achieves the best matching result, and in other occlusion scenes with deeper occlusion degrees, patch 1×3 achieves a better matching result. In addition, the greater the inclination of the polyline, the worse the corresponding anti-occlusion performance. As shown in Fig. 22, the polyline corresponding to patch 1×3 has the smallest inclination angle, so its anti-occlusion performance is the best.

C. ALGORITHM PERFORMANCE COMPARISON

In this section, we will conduct Experiment 8 to demonstrate the anti-occlusion performance of the proposed MSTMM scheme by comparing the performance of several algorithms. According to the results of Experiment 1 to Experiment 7, the configurations of the MSTMM scheme involved in the comparison for this experiment are as follows. The image patch

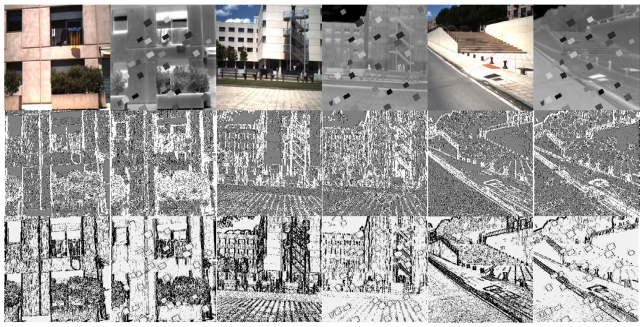


FIGURE 24. The examples of mapped images on the test set. From top to bottom are: input visible and thermal infrared images, mapped images generated by the traditional special integer mapping mechanism in MSTMM-IM, mapped images generated by the neural network mapping mechanism in MSTMM-NM.

used is 1×3, the mapping mechanisms used are MSTMM-IM and MSTMM-NM. The interval size of the visible images in the training set will be set to eight. The training set and test set used in this experiment are the same as in Experiment 4. In the MSTMM-IM mechanism, the distance threshold of patch 1×3 is four for visible template images and two for thermal infrared query images. In the MSTMM-NM mechanism, the distance threshold of patch 1×3 is five for visible template images and five for thermal infrared query images. As in Fig. 23, show the statistical results of the success rate for this experiment with different algorithms under four different occlusion levels.

The results in Fig. 23 show that the proposed MSTMM scheme achieves the best success rate compared with NCC, MTM, STVO, LCTS, M-RGBIR, MUNIT, and NCCNet algorithms. In addition, the greater the inclination of the polyline, the worse the corresponding anti-occlusion performance. As shown in Fig. 23, among the top five algorithms, the polyline corresponding to MSTMM-NM has the smallest inclination angle, so its anti-occlusion performance is the best, especially compared with the STVO algorithm with complex scene matching ability. Note that MUNIT here refers to the process of using the MUNIT algorithm to generate the transformation map of the visible image, then replacing the original visible image with the transformation map, and then using the NCC algorithm to achieve template matching. The performance of the MSTMM-NM mechanism based on a neural network is better than the traditional MSTMM-IM mechanism, which shows that the MSTMM scheme proposed in this paper has obvious advantages whether it is the traditional mechanism or the mechanism based on a neural network. In Fig. 23, the MUNIT statistical results are obtained under the assumption that the model converges when trained for 1000 epochs, and the M-RGBIR statistical results are obtained under the assumption that the model converges when trained for 30 epochs, and the NCCNet statistical results are obtained under the assumption that the model converges when the network training loss does not improve for ten consecutive epochs. In addition, the low performance of the

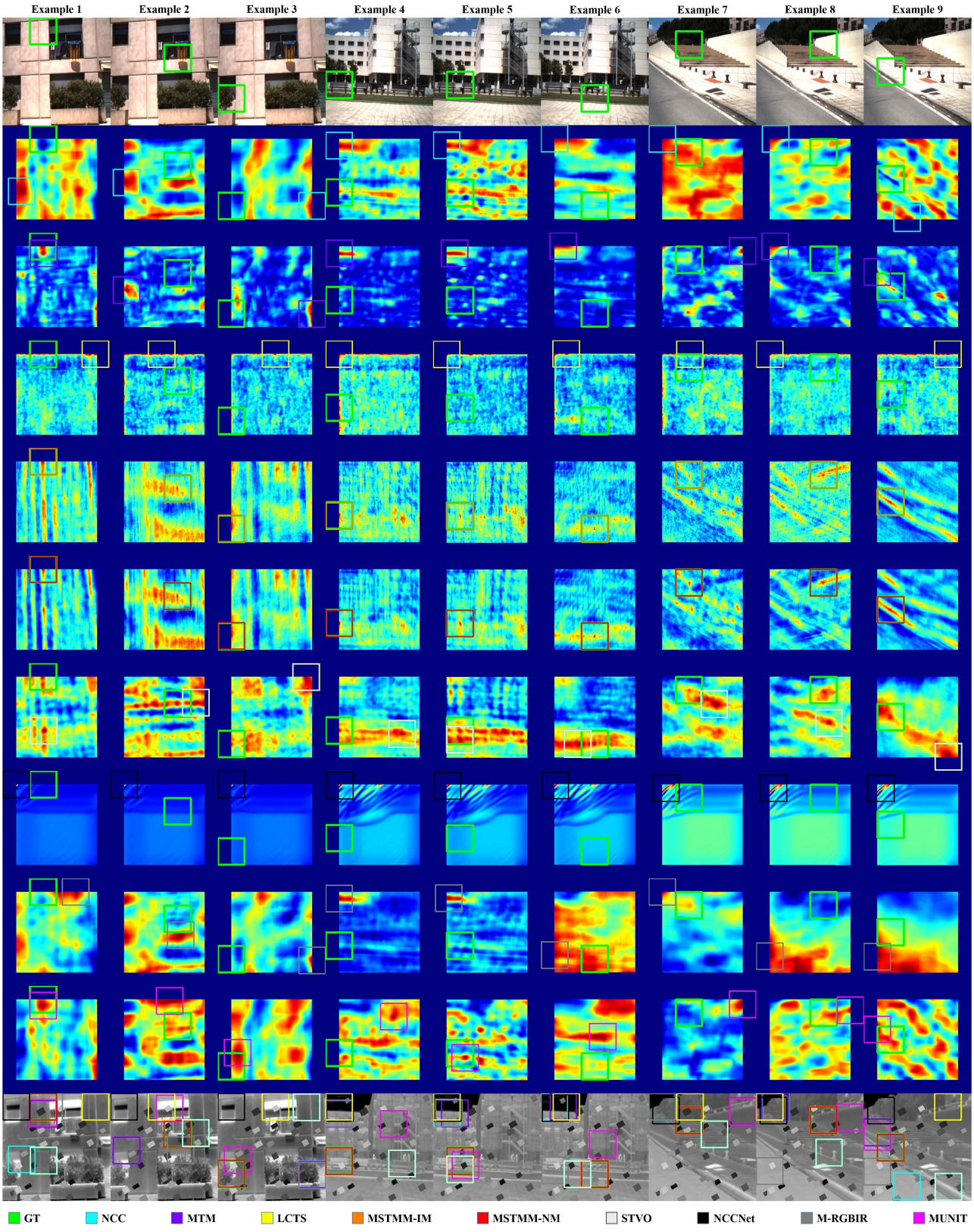


FIGURE 25. The nine examples of detection results for nine different algorithms on the test set. From top to bottom are: template image, hot maps (from top to bottom are the hot maps of nine algorithms of NCC, MTM, LCTS, MSTMM-IM, MSTMM-NM, STVO, NCCNet, M-RGBIR, and MUNIT), query image. The template image is marked in green over the visible images; Each hot map is marked with the groundtruth (GT) and detection result, i.e., its ideal position and actual global maximum position; The query image is marked with the groundtruth (GT) and the detection results of nine different algorithms. See the legend at the bottom for the representative colors of groundtruth (GT) and nine different algorithms.

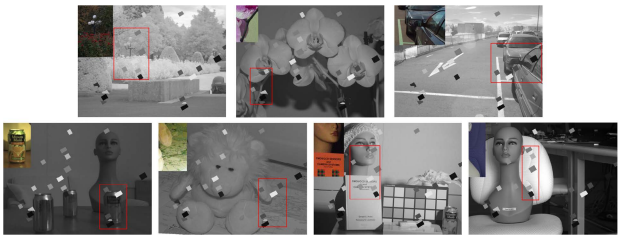


FIGURE 26. The seven examples of the template matching results of the MSTMM-NM scheme on the RGB-NIR dataset. The upper left corner of the images is the visible template image, and the red box is the matching bounding box.

NCCNet algorithm in Fig. 23 is caused by too few training sets after filtering out unqualified data. We will demonstrate in Experiment 10 that the MSTMM-NM mechanism in this paper outperforms the MUNIT, M-RGBIR, and NCCNet algorithms regardless of the number of epochs required for the model to converge or the time required to train an epoch.

A quantitative comparison of the MSTMM scheme is given in Fig. 23, and a qualitative comparison is shown in Fig. 25. Note that the query images in Fig. 25 are thermal infrared images under occlusion levels of three. The MSTMM scheme requires that the input image be mapped to a map image by either a traditional special integer mapping mechanism in MSTMM-IM or a neural network mapping mechanism in MSTMM-NM before template matching. The examples of mapped images are shown in Fig. 24. In Fig. 24, subjectively, the neural network mapping mechanism in MSTMM-NM is better than the traditional special integer mapping mechanism in MSTMM-IM. In Fig. 25, the MSTMM scheme successfully achieves template matching in almost all of these examples. In the hotmap of Fig. 25, the contrasting algorithms show low distinction, but the MSTMM scheme concentrates on the surrounding region of interest, showing a distinct mode around the groundtruth location.

D. ADDITIONAL EXPERIMENT: ALGORITHM PERFORMANCE ON RGB-NIR

In order to better evaluate the performance of the MSTMM-NM scheme in this paper, the qualitative evaluation results of the MSTMM-NM scheme on the RGB-NIR (RGB and near-infrared) dataset [51] are shown in Fig. 26. The MSTMM-NM scheme successfully achieves template matching in all of these examples.

E. RUNTIME ANALYSIS AND COMPARISON

The MSTMM scheme runtime test experiment is divided into two main experiments: the first experiment is to test the runtime of the scheme and compare it with other algorithms; the second experiment is to test the runtime required for the network to train an epoch and the number of epochs required to train until the model converges.

1) EXPERIMENT 9: RUNTIME

The runtime of the sliding window template matching algorithm is mainly consumed in the window sliding process, so this experiment mainly tests the runtime of the MSTMM scheme with different sizes of query images and different sizes of template images. The size of the template image used for the different-sized query images experiments is 64×64, and the query image sizes range from 128×128 to 704×704 with an interval of 64, with a total of ten different sizes. The size of the query image used for the different-sized template images experiments is 512×512, and the size of the template images size ranges from 64×64 to 352×352 with an interval of 32, with a total of ten different sizes. The configuration of the MSTMM scheme in this experiment is the same as in Experiment 8. As in Fig. 27, show the average runtime statistics of the 10 experiments for the different algorithms.

The results in Fig. 27 show that the proposed MSTMM scheme achieves almost the highest computational efficiency compared with STVO, NCCNet, MTM, and MUNIT algorithms, except for the NCC algorithm that uses the extremely efficient OpenCV library function. The MUNIT statistical results in Fig. 27 exclude model loading time. Since the MSTMM-NM model has only a limited number of weights, and the weights are only related to the size of the mapped value, the trained model can directly extract these weights into a ''.txt'' file, and then use these weights in the imported ''.txt'' file for template matching to save the model loading time. Note that since the LCTS and M-RGBIR algorithms are too inefficient relative to other algorithms, this experiment does not present the runtime polylines of these two algorithms.

2) EXPERIMENT 10: TIME REQUIRED IN TRAINING

In this experiment, the information statistics on the training process of M-RGBIR, MSTMM-NM, NCCNet, and MUNIT algorithms in Experiment 8 are carried out, and the statistical results are shown in Fig. 28. For NCCNet, the training statistical results are obtained under the assumption that the model converges when the network training loss does not improve for ten consecutive epochs; for MSTMM-NM, the training statistical results are obtained when the training accuracy and loss are stable within ten epochs, and the training set loading time is not within the time statistics; for MUNIT, the training statistical results are obtained under the assumption that the model converges when trained for 1000 epochs; for M-RGBIR, the statistical results are obtained under the assumption that the model converges when trained for 30 epochs.

As shown in Fig. 28, the MSTMM-NM mechanism proposed in this paper can complete training in as few as 58 epochs, multiplying the time required to train one epoch is about 1.8 seconds, and the total training time is about 104.4 seconds. Compared with M-RGBIR, NCCNet, and MUNIT algorithms, the training time of the MSTMM-NM mechanism is significantly reduced, allowing for online training. Furthermore, since this experiment is performed with a very limited device platform configuration and only uses

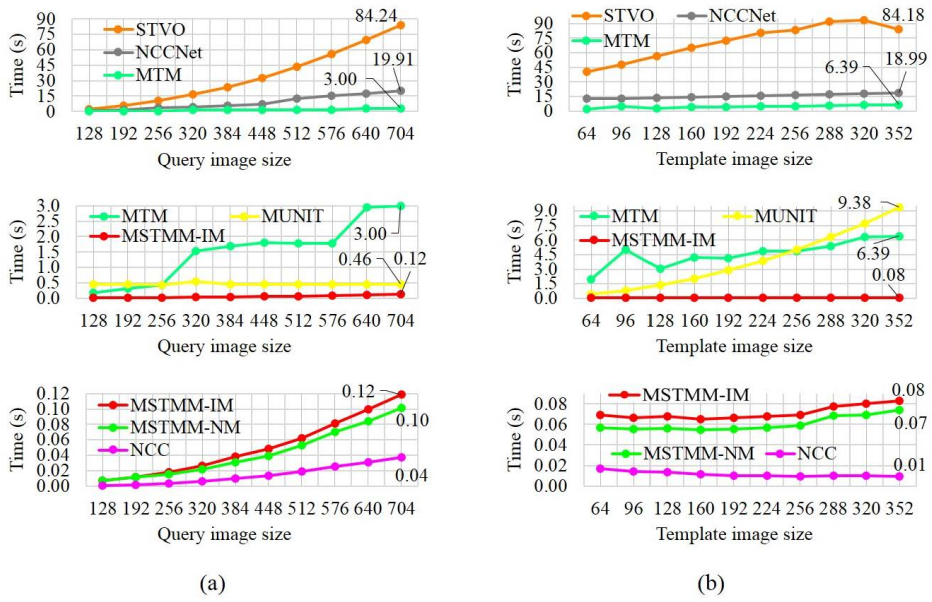


FIGURE 27. The runtime of different algorithms (see legend) with different sizes of query images and different sizes of template images: (a) comparison of seven algorithms of STVO, NCCNet, MTM, MUNIT, MSTMM-IM, MSTMM-NM, and NCC under different sizes of query images; (b) comparison of seven algorithms of STVO, NCCNet, MTM, MUNIT, MSTMM-IM, MSTMM-NM, and NCC under different sizes of template images. Since the sizes of the two dimensions of the query image in (a) and the sizes of the two dimensions of the template image in (b) are equal, the horizontal axis in the figure only shows the size of one dimension. The tag value is the runtime when the query image size is 704 × 704 in (a) and the template image size is 352 × 352 in (b). The runtime of various algorithms can be shown in one figure, but for more clear showing, we use three figures to present the experimental results. The runtime polyline of the MTM algorithm in the first and second figures in (a) and (b) respectively are the same. The runtime polyline of the MSTMM-IM algorithm in the second and third figures in (a) and (b) respectively are the same.

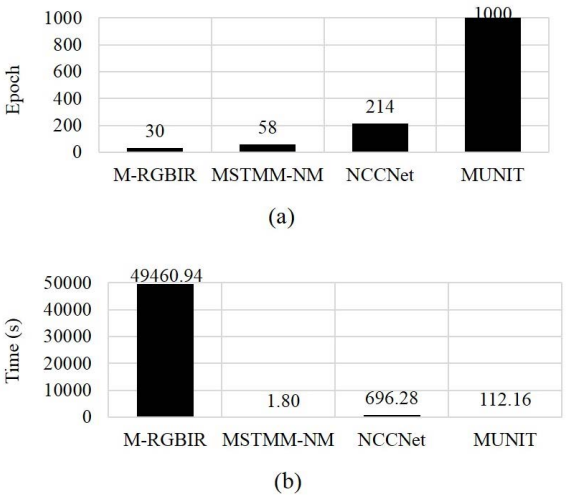


FIGURE 28. The training process information for different algorithms: (a) the minimum number of training epochs; (b) the runtime required for one training epoch. The tag value is the training information corresponding to the algorithm.

CPUs instead of GPUs for training, it is entirely possible to train the network online using idle time.

V. CONCLUSION

In this paper, a visible and thermal infrared images template matching scheme under occluded scenarios, called Matching by Slice Transform Matrix Mapping (MSTMM), is proposed to address the issue of efficiency and robustness of matching in the case of limited hardware platform resources. The Expanded Slice Transform with Adaptive Gray Level (EST-AGL) matrix developed for the MSTMM scheme can effectively extract information about the distribution of pixels with the same gray level, which can be used as the structural feature of heterogeneous images with nonlinear intensity differences, and then the MSTMM scheme maps the EST-AGL matrices corresponding to different modality images into correlation surface images with the same modality through the traditional integer mapping mechanism or neural network mapping mechanism.

The three flexible mapping implementation methods, the traditional integer mapping method, the offline neural network training method, and the online neural network training method, make it possible for the scheme to be successfully implemented on different hardware platforms. In neural network training, the usage of a minimalist fully connected FNN rather than the popular CNN brings the possibility of online training. Since the MSTMM-NM network model has only a limited number of weights, and the weights are only related to the size of the mapped value, the trained model can directly extract these weights into a ".txt" file, and then use these weights in the imported ".txt" file for template matching. The advantage of this is that the trained model can handle input

images of arbitrary size, avoiding the problem of most deep learning algorithms accepting only fixed-size input images for inference.

Finally, the experimental results show that the MSTMM scheme outperforms many existing popular algorithms in terms of efficiency and robustness. According to the experiment, properly filtering the training set of MSTMM-NM, and reducing the number of negative samples in the training set has the potential to improve the performance of MSTMM-NM, and in the future, we will explore the performance of the algorithm when the number of negative samples in the training set decreases.

REFERENCES

[1] W. Zhang, X. Sui, G. Gu, Q. Chen, and H. Cao, "Infrared thermal imaging super-resolution via multiscale spatio-temporal feature fusion network," *IEEE Sensors J.*, vol. 21, no. 17, pp. 19176–19185, Sep. 2021.

[2] S. Cui, A. Ma, Y. Wan, Y. Zhong, B. Luo, and M. Xu, "Cross-modality image matching network with modality-invariant feature representation for airborne-ground thermal infrared and visible datasets," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[3] L. Liu, M. Chen, M. Xu, and X. Li, "Two-stream network for infrared and visible images fusion," *Neurocomputing*, vol. 460, pp. 50–58, Oct. 2021.

[4] Y. Zhao, G. Fu, H. Wang, and S. Zhang, "The fusion of unmatched infrared and visible images based on generative adversarial networks," *Math. Problems Eng.*, vol. 2020, pp. 1–12, Mar. 2020.

[5] X. Qian, M. Zhang, and F. Zhang, "Sparse GANs for thermal infrared image generation from optical image," *IEEE Access*, vol. 8, pp. 180124–180132, 2020.

[6] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[7] I. Ullah, F. Yang, R. Khan, L. Liu, H. Yang, B. Gao, and K. Sun, "Predictive maintenance of power substation equipment by infrared thermography using a machine-learning approach," *Energies*, vol. 10, no. 12, p. 1987, Dec. 2017.

[8] Z. Xu, J. Wang, P. Xu, and T. Liu, "Infrared image temperature measurement based on FCN and residual network," in *Proc. Chin. Intell. Syst. Conf. (CISC)*. Singapore: Springer, 2017, pp. 769–775.

[9] Y. Nakayama, G. Sun, S. Abe, and T. Matsui, "Non-contact measurement of respiratory and heart rates using a CMOS camera-equipped infrared camera for prompt infection screening at airport quarantine stations," in *Proc. IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. (CIVEMSA)*, Jun. 2015, pp. 1–4.

[10] Y. Chen, F. Huang, B. Liu, Z. Wang, S. Zhang, and H. Zhong, "Feasibility analysis of ultra-wide FOV infrared imaging system applied in vehicle auxiliary driving," in *Proc. 6th Symp. Novel Optoelectronic Detection Technol. Appl.*, Apr. 2020, pp. 792–801.

[11] D.-G. Gwak and D. H. Kim, "Image matching algorithm for thermal panorama image construction adaptable for fire disasters," *J. Inst. Control, Robot. Syst.*, vol. 22, no. 11, pp. 895–903, Nov. 2016.

[12] S. K. M. Kamath, K. Panetta, and S. Agaian, "Multi-view near-infrared image mosaicking for face detection in smart cities," in *Proc. SPIE*, vol. 10668, pp. 121–129, May 2018.

[13] J. Lu, M. Hu, J. Dong, S. Han, and A. Su, "A novel dense descriptor based on structure tensor voting for multi-modal image matching," *Chin. J. Aeronaut.*, vol. 33, no. 9, pp. 2408–2419, Sep. 2020.

[14] J. M. Kim, J.-M. Park, and J.-W. Lee, "Comparison between traditional and CNN based stereo matching algorithms," *J. Inst. Control, Robot. Syst.*, vol. 26, no. 5, pp. 335–341, May 2020.

[15] K. Nan, H. Qi, and Y. Ye, "A template matching method of multi-modal remote sensing images based on deep convolutional feature representation," *Acta Geodaetica Et Cartographica Sinica*, vol. 48, no. 6, pp. 727–736, 2019.

[16] H. Zhang, N. Shao, X. Meng, and A. Wang, "Fast image matching method and its applications in underwater positioning," in *Proc. Int. Conf. Electr. Control Eng.*, Wuhan, China, Jun. 2010, pp. 970–973.

[17] S. K. Sahani and M. S. Chauhan, "A novel fast template matching algorithm based on bounded approach using block partitioning method," in *Proc. 2nd Int. Conf. Range Technol. (ICORT)*, Aug. 2021, pp. 1–6.

[18] R. Zhang, C. Mu, X. Gao, K. Liu, and Y. Ma, "A fusion algorithm of template matching based on infrared simulation image," in *Proc. SPIE*, vol. 10033, pp. 30–34, Aug. 2016.

[19] J. N. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *Proc. Int. Conf. Adv. Comput., Control, Telecommun. Technol.*, Trivandrum, Kerala, Dec. 2009, pp. 819–822.

[20] Y. Li, J. Chen, M. Ke, L. Li, Z. Ding, and Y. Wang, "Small targets recognition in SAR ship image based on improved SSD," in *Proc. IEEE Int. Conf. Signal, Inf. Data Process. (ICSIDP)*, Dec. 2019, pp. 1–6.

[21] M. Srikham, C. Pluempitiwiriyawej, and T. Chanwimaluang, "Comparison of dense matching algorithms in noisy image," in *Proc. SPIE*, vol. 7546, pp. 207–220, Feb. 2010.

[22] X. Li, Y. Hu, T. Shen, S. Zhang, J. Cao, and Q. Hao, "A comparative study of several template matching algorithms oriented to visual navigation," in *Proc. SPIE*, vol. 11550, pp. 66–74, Oct. 2020.

[23] X. Xiong, Q. Xu, G. Jin, H. Zhang, and X. Gao, "Rank-based local self-similarity descriptor for optical-to-SAR image matching," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1742–1746, Oct. 2020.

[24] Q. Hou, C. Bian, L. Lu, and W. Zhang, "Template matching based on weighted voting accumulation measure," *Guangxue Jishu/Opt. Tech.*, vol. 39, no. 1, pp. 23–27, 2013.

[25] S. Koley, P. K. Dutta, and I. Aganj, "Radius-optimized efficient template matching for lesion detection from brain images," *Sci. Rep.*, vol. 11, no. 1, pp. 1–21, Dec. 2021.

[26] X. Jiang, J. Ma, G. Xiao, Z. Shao, and X. Guo, "A review of multimodal image matching: Methods and applications," *Inf. Fusion*, vol. 73, pp. 22–71, Sep. 2021.

[27] J. Cheng, Y. Wu, W. Abdalmageed, and P. Natarajan, "QATM: Quality-aware template matching for deep learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11545–11554.

[28] J. Cheng, Y. Wu, W. Abd-Almageed, and P. Natarajan, "Image-to-GPS verification through a bottom-up pattern matching network," in *Proc. 26th IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 546–561.

[29] G. Kovacs, "Matching by monotonic tone mapping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1424–1436, Jun. 2018.

[30] A. Mahmood and S. Khan, "Correlation-coefficient-based fast template matching through partial elimination," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2099–2108, Apr. 2012.

[31] Y. Yaguchi, K. Iseki, and R. Oka, "Full pixel matching between images for non-linear registration of objects," *IPSJ Trans. Comput. Vis. Appl.*, vol. 2, pp. 1–14, May 2010.

[32] Y. Ye, J. Shan, L. Bruzzone, and L. Shen, "Robust registration of multimodal remote sensing images based on structural similarity," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2941–2958, Mar. 2017.

[33] Q. Wu, G. Xu, Y. Cheng, Z. Wang, W. Dong, and L. Ma, "Robust and efficient multi-source image matching method based on best-buddies similarity measure," *Infr. Phys. Technol.*, vol. 101, pp. 88–95, Sep. 2019.

[34] D. Buniatyan, S. Popovych, D. Ih, T. Macrina, J. Zung, and H. S. Seung, "Weakly supervised deep metric learning for template matching," in *Advances in Computer Vision* (Advances in Intelligent Systems and Computing). Cham, Switzerland: Springer 2020, pp. 39–58.

[35] R. Zhu, D. Yu, S. Ji, and M. Lu, "Matching RGB and infrared remote sensing images with densely-connected convolutional neural networks," *Remote Sens.*, vol. 11, no. 23, p. 2836, Nov. 2019.

[36] L. Hu and Y. Zhang, "Facial image translation in short-wavelength infrared and visible light based on generative adversarial network," *Acta Optica Sinica*, vol. 40, no. 5, 2020, Art. no. 0510001.

[37] F. Wu, W. You, J. S. Smith, W. Lu, and B. Zhang, "Image-image translation to enhance near infrared face recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3442–3446.

[38] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[39] Y. Luo, D. Pi, Y. Pan, L. Xie, W. Yu, and Y. Liu, "ClawGAN: Claw connection-based generative adversarial networks for facial image translation in thermal to RGB visible light," *Exp. Syst. Appl.*, vol. 191, Apr. 2022, Art. no. 116269.

[40] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 179–196.

[41] H. Jin and L. L. Zou, "Detection of hidden disease of concrete bridge based on infrared thermal imaging," *J. Phys. Conf.*, vol. 1748, no. 4, 2021, Art. no. 042041.

[42] T. Dekel, S. Oron, M. Rubinstein, S. Avidan, and W. T. Freeman, "Best-buddies similarity for robust template matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2021–2029.
[43] I. Talmi, R. Mechrez, and L. Zelnik-Manor, "Template matching with deformable diversity similarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1311–1319.
[44] S. Korman, S. Soatto, and M. Milam, "OATM: Occlusion aware template matching by consensus set maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2675–2683.
[45] Q. Ren, Y. Zheng, P. Sun, W. Xu, D. Zhu, and D. Yang, "A robust and accurate end-to-end template matching method based on the Siamese network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[46] A. Toet. (Apr. 2014). *TNO Image Fusion Dataset*. Figshare. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029/1
[47] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 12484–12491.
[48] Y. Hel-Or, H. Hel-Or, and E. David, "Matching by tone mapping: Photometric invariant template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 317–330, Feb. 2014.
[49] Y. Ye and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *ISPRS J. Photogramm. Remote Sens.*, vol. 90, pp. 83–95, Apr. 2014.
[50] C. A. Aguilera, A. D. Sappa, and R. Toledo, "LGHD: A feature descriptor for matching across non-linear intensity variations," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 178–181.
[51] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multi-spectral registration for natural images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 309–324.

LICHUN MEI received the B.Eng. degree in electronic information science and technology from Luoyang Normal University, Henan, China, in 2013, and the M.S.E. degree in electronics science and technology from the PLA University of Science and Technology, Jiangsu, China, in 2017. He is currently pursuing the Ph.D. degree in communication and information system with the College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu. His research interests include image matching, heterologous image processing, computer vision, artificial intelligence, data analysis, and industrial control.

HUAIYE WANG received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China, in 2005. He is currently an Engineer at Beijing Space Feiteng Equipment Technology Company Ltd., China. His research interests include navigation, guidance, and control.

CAIYUN WANG received the Ph.D. degree in signal and information processing from the Beijing University of Aeronautics and Astronautics, China, in 2007. She is currently a Professor with the Nanjing University of Aeronautics and Astronautics, China. Her research interests include computer vision, artificial intelligence, radar signal processing, and radar target detection.

YUANFU ZHAO (Senior Member, IEEE) received the Ph.D. degree from the Shaanxi Microelectronics Technology Institute, China, in 1989. He is currently a Professor with the Beijing Microelectronic Technology Institute. His main research interests include radiation effects and hardening of devices and integrated circuits.

JUN ZHANG received the M.S. degree from the Beijing University of Aeronautics and Astronautics, China. He is currently an Engineer at Beijing Space Feiteng Equipment Technology Company Ltd., China. His research interests include navigation, guidance, and control.

XIAOXIA ZHAO received the Ph.D. degree from the China University of Mining and Technology, Beijing, China. She is currently an Engineer at Beijing Space Feiteng Equipment Technology Company Ltd., China. Her research interests include digital image processing and automatic target recognition.