IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 July 2023, accepted 23 July 2023, date of publication 27 July 2023, date of current version 2 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3299266

RESEARCH ARTICLE

Research on Infrared and Visible Image Registration Algorithm for Complex Road Scenes

YUAN WANG, XIANGYANG LIANG, AND LEI CHEN
School of Computer Science and Technology, Xi'an Technological University, Xi'an 710000, China
Corresponding author: Xiangyang Liang (xiangyangl0913@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62276146.

ABSTRACT This study proposes a novel image registration algorithm to solve the problem of low registration accuracy caused by excessive difference in resolution and spectral differences between infrared and visible images. First, we use a VI-CycleGAN network to translate visible images into fake infrared images. Furthermore, we incorporate a normalization-based attention mechanism (NAM) into each residual block to capture the global information of the image and preserve its details. Meanwhile, we consider higher-level semantic information and introduce a hybrid loss function that better preserves the content features of the original image. We then use guided filtering to process the infrared image and fake infrared image to eliminate complex background noise. Subsequently, coarse registration was performed using the Speeded-Up Robust Features (SURF) algorithm. Finally, we used the random sample consensus (RANSAC) algorithm to remove the mismatch points and achieve accurate registration. This study conducted experiments on two different VIS-IR image datasets and compared Gaussian field estimator with manifold regularization (GFEMR), radiation-variation insensitive feature transform (RIFT), and locally normalized image feature transform (LNIFT) algorithms. The experimental results show that compared with LNIFT algorithms, the registration accuracy of the proposed method was improved by approximately 5%.

INDEX TERMS Image registration, feature matching, infrared image, visible image, feature descriptor.

I. INTRODUCTION

Fusion of visible and infrared images is widely used in computer vision applications to provide a more comprehensive representation of a scene. Infrared images capture thermal radiation information, whereas visible images offer high resolutions [1]. The Integration of these two types of images requires registration technology. Image registration is a preprocessing stage for image fusion, which directly affects its effectiveness of image fusion. Image registration is also essential for object localization, 3D reconstruction, cross-modal pedestrian recognition, image fusion, and image stitching. Therefore, there is significant value in conducting research on algorithms for registering infrared and visible images.

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

Existing methods for registering heterogeneous images can be classified into two categories: intensity-based and feature-based [2]. The intensity approach matches images by utilizing similarity metrics, such as normalized cross-correlation (NCC) [3] and mutual information (MI) [4]. Although this method is effective for image registration within a single mode, it cannot solve the problem of low registration accuracy owing to the significant grayscale variations. The feature-based approach is currently the most common method used for image registration. It involves extracting prominent features from images, such as those identified by the Harris algorithm, Scale-Invariant Feature Transform (SIFT) [5], Speeded Up Robust Features (SURF), Canny edge detection algorithm [6], and morphological methods, and uses them as reference information for matching two images. Feature-based methods can effectively solve the homologous image registration problem. Owing to the significant

nonlinear intensity differences between heterogeneous images, traditional image registration methods are not directly applicable to infrared and visible images.

To solve the aforementioned problem, this article uses generative adversarial networks to convert heterogeneous images to their corresponding homogeneous images. GAN [7] has been effective in image styles transfer tasks, such as DualGAN [8], SpaGAN [9], CycleGAN [10], and others.

Therefore, this study proposes a registration method based on the VI-CycleGAN network combined with guided filtering to transform heterogeneous image registration into a homogeneous registration problem. First, the attention mechanism and mixed loss were incorporated into the CycleGAN network for optimization. The NAM attention mechanism calculates the attention weights based on global information, which helps the VI-CycleGAN network better learn the mapping relationship between images. The mixed loss helps the VI-CycleGAN generator to better preserve the edge information and feature representation of the original and real images, thereby improving the quality and realism of the generated images. Then, we used the Guided Filter algorithm to denoise and enhance the generated fake infrared images and infrared images. The SURF algorithm was used for coarse matching. The SURF algorithm is a robust and efficient feature detection and extraction method that can be used to achieve the accurate registration of homologous images. Finally, we used the RANSAC algorithm to eliminate mismatching points and achieve precise registration.

The main contributions of this paper are as follows:

(1) This study proposes an infrared and visible registration algorithm based on the VI-CycleGAN, which reduces the effect of different representations of multimodal images.
(2) This study optimizes the CycleGAN style translation network by introducing the NAM attention mechanism and mixed loss function.
(3) This study used the guided filtering method to solve the problem of difficult registration caused by excessive complex background noise.

The rest of the paper is organized as follows. Section II presents related work. Section III presents the details of the proposed method and VI-CycleGAN network. Section IV describes the image denoising and alignment process. Section V presents the experimental setup and results. Finally, Section VI concludes the paper.

II. RELATED WORK

Combining images from multiple sensors through information fusion can yield a wealth of information, and matching data from different sensors can provide a more reliable interpretation of the image regions or specific targets. However, matching can be challenging because of the significant differences between the sensors. Therefore, most studies have used feature matching methods to solve these problems. Typical feature-based matching methods include image preprocessing, feature detection, feature description, transform modeling, and feature matching [11]. The detected features can be classified as corner points, edge contours, and area features. Point features are more common than line and area features in image matching.

Li et al. [12] proposed a method for registering infrared and visible images based on constrained point features to address the high complexity of traditional point features. This approach avoids the construction of complex feature descriptors and introduces advanced semantic information to improve the registration accuracy. Paul et al. [13] proposed an improved version of the traditional SIFT algorithm to increase the number of matching features between the images. Gao et al. [14] developed a registration method for multisource remote sensing images based on multi-scale feature point matching, utilizing an improved Harris corner detection algorithm and the PIIFD feature descriptor, which achieved accurate registration. Wang et al. [15] proposed a popular regularization Gaussian field estimator combined with the SURF-PIIFD algorithm to address the problem of registering different modalities of retinas, and demonstrated the effectiveness of this method through extensive experiments. Li et al. [16] proposed to use temporal consistency to solve the problem of radiative and geometric transformations while detecting only corner and edge points with better robustness. Li et al. [17] proposed a spatial domain multimodal feature matching algorithm to solve the problem of severe nonlinear radiation distortion. Jiang et al. [18] introduced a contour angle orientation CAO-C2F algorithm, which utilizes the CSS algorithm to compute the main orientation of contour points and combine it with the SIFT descriptor to achieve precise registration. These methods are based on feature point matching and mainly improve feature point extraction methods to extract more robust features. However, such methods cannot effectively overcome the noise in complex backgrounds and extract many invalid features, which largely affects the feature point extraction in complex road scenes.

Wang et al. [19] proposed an improved feature detection and description method based on image edge structures to enhance the matching repeatability and accuracy of infrared and visible image registration. The proposed method was tested on two datasets, and showed an improved registration accuracy. Zhao [20] addressed the problems of time consumption and matching accuracy, and proposed a registration method that combined local edge information with SURF feature points. Legg et al. [21] proposed a feature neighborhood mutual information similarity measurement method based on mutual information (MI) to achieve good results in the registration of multimodal retinal fundus images. Ma et al. [22] proposed a local linear transformation (LLT) to estimate the transformation model and address the problem of large matching outliers. The proposed method was evaluated on a multimodal remote sensing dataset and yielded accurate registration results. The above method is based on matching with fused features, which also improves the feature extraction method to enable the key points to focus more on

useful features. However, because of the significant differences between infrared and visible images, whether observed from the intensity, edge, or texture, the differences were significant. Therefore, such methods cannot effectively overcome the large differences between different spectra.

To address the problems of heterogeneous image alignment for complex scenes, this study uses a modal transformation approach to improve existing algorithms. However, in the era of deep learning, owing to the inherent challenges in addressing multimodal image-matching variability and geometric deformation, a growing number of methods and diversity have been proposed. Huang et al. [23] proposed an unsupervised image-to-image translation framework, MUNIT, for multimodal image translation between real image domains. Isola et al. [24] proposed a conditional generative adversarial network (cGAN) to learn mapping between pairs of images, which has been widely used in the field of image generation. Zhu et al. [10] proposed a method for learning to transform images from the source domain X to the target domain Y without paired examples.

The method proposed in this paper is based on the CycleGAN style transfer network, and introduces an NAM attention mechanism and guided filtering algorithm. The loss function of the network was also improved. The method utilizes the SURF algorithm and RANSAC for feature point matching, and the experimental results show the improved effectiveness of modality conversion and matching accuracy.

III. PROPOSED METHOD

This section introduces the proposed method for image registration, which involves a five-step process: multimodal translation, guided filtering processing, feature extraction and feature matching, and outlier removal. A flowchart of the proposed method is shown in **Figure 1**. Each module is briefly described as follows.

As shown in **Figure 1**. First, because of the significant representation differences between visible and infrared images, this study used a VI-CycleGAN network to translate visible images into approximate infrared images. We define the transformed image as a fake infrared image. Simultaneously, a NAM attention mechanism is incorporated into the network to better capture the global information in the images and retain the image details. The traditional mean squared error is used to define the loss function in the network. This study considers higher-level semantic information and introduces a mixed loss to better preserve the content features of the original images. Next, because noise and artifacts are introduced in the image translation, guided filtering is introduced for noise removal and smoothing. At this point, we need to carry out image feature extraction and feature matching, which can obtain the key information of the image. Therefore, this study uses the SURF algorithm to extract feature points and feature point alignment for the image. Finally, we used the RANSAC algorithm for false match elimination to obtain the final alignment results.

A. VI-CYCLEGAN NETWORK

In this study, we propose an improved version of the VI-CycleGAN neural network model for style transfer. Our approach aims to capture the essential features of images. These features include pedestrian brightness and the background edge structure. We achieved this by integrating a Normalization-based attention mechanism called NAM [25] into the CycleGAN model. The NAM mechanism was designed to focus on these salient details during the image translation process. Furthermore, we introduce edge loss and perceptual loss as additional optimization targets to enhance the transformation effect. This is done to further enhance the transformation strength of the thermal targets and improve the edge contours of the background. The generator, discriminator, and loss function of the VI-CycleGAN model are shown in **Figure 2.** The data domains X and Y are visible and infrared images, respectively. the image in the X domain is generated as a Y domain image by G and reconstructed back to the X domain by F; the image in the Y domain is generated as an X domain image by generator F and reconstructed as a Y domain image by the G generator. The discriminator $D(X)$ distinguishes between X domain images and $F(Y)$, while the discriminator $D(Y)$ distinguishes between Y domain images and $G(X)$ images.

B. ATTENTION MECHANISM

The attention mechanism highlights important features and suppresses irrelevant information by applying different weights to image features. As part of the multimodal image transformation process, the network is required to learn different feature mappings. For example, infrared images need to focus on the thermal target intensity and edge contour information, whereas visible images need to prioritize image detail information. Attention mechanisms include spatial domain, channel domain [26], and mixed attention mechanisms.

NAM is an efficient and lightweight attention mechanism that integrates module sets from CBAM [27]. It improves network performance by redesigning channel and spatial attention sub-modules and using standard deviation to represent weight importance. To enhance image feature analysis, this study introduced a normalization-based attention module (NAM) into the residual blocks of the VI-CycleGAN generator network. The NAM attention mechanism was embedded in each ResBlock of both G and F generators in VI-CycleGAN. **Figure 3** illustrates the generator's structural composition in the VI-CycleGAN network, where Conv, ResNet, and Deconv represent the convolution layer, residual block, and deconvolution layer, respectively. Generator G includes an encoder, transformer, and decoder, with the transformer containing nine residual blocks that maintain the same input and output sizes, allowing for adaptive network depth and reducing gradient vanishing.

Our experiments demonstrated that the integration of NAM into residual blocks improves the overall performance of the

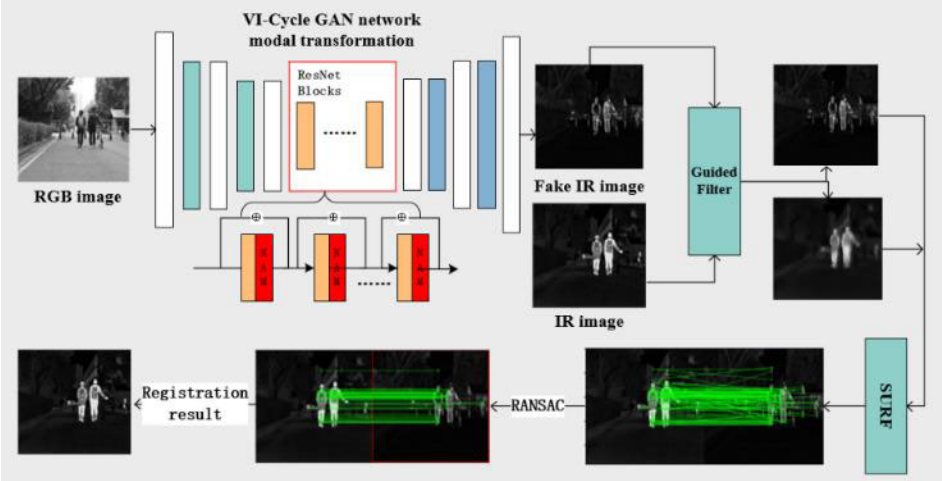

FIGURE 1. Flowchart of the proposed registration method.

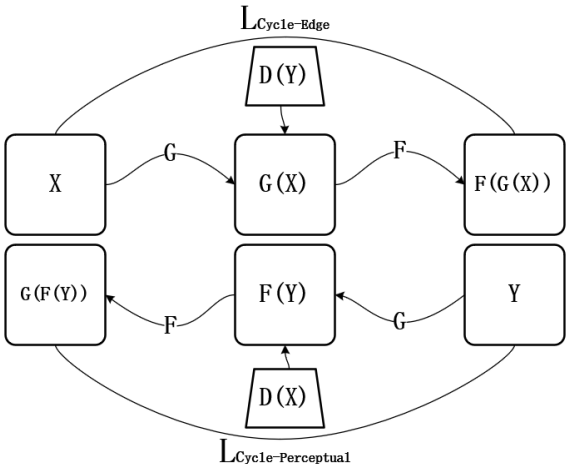

FIGURE 2. VI-CycleGAN network model.

network. The NAM mechanism provides a means to selectively weight feature maps, thereby emphasizing the most informative features while suppressing noise and irrelevant information. This leads to more efficient and effective feature extraction and ultimately better network performance.

C. LOSS FUNCTION

The loss function of the proposed algorithm includes generative adversarial loss, cycle consistency loss, Cycle-Edge loss [28],and Cycle-Perceptual loss [29] to generate more realistic fake infrared images [30]. The generative adversarial loss function is the same as the CycleGAN loss function. The VI-CycleGAN cycle consistency loss learns both the G and F mappings and aims to make $F(G(x))$ similar to x. As the prominent feature of visible images is the contour edge feature, edge loss is introduced in this study to maintain the Cycle-Edge consistency of the generated images. This can induce the generator to better retain the edge information between visible images, thus making the generated IR images more realistic and accurate.

The introduction of the Cycle-Edge loss function can significantly improve the quality and accuracy of image translation compared with using only the traditional mean square error loss function. The Cycle-Perceptual loss can obtain more information than the MSE loss by computing the space transformed into the feature space. The introduction of a Cycle-Perceptual loss function can better preserve the content features and semantic information of the images, which can generate more realistic and natural-looking IR images. The Cycle-Edge loss is defined as follows:

$$L_{Cycle-Edge} = \|\nabla F(G(X)) - \nabla(X)\|_1 + \|\nabla G(F(Y)) - \nabla(Y)\|_1 \quad (1)$$

The ∇ in the formula refers to the Laplace operator, X and Y represent the visible image and infrared image respectively. The Cycle-Perceptual loss is defined as follows:

$$L_{Cycle-Perceptual} = \|\psi jF(G(X)) - \psi j(X)\|_2^2 + \|\psi jG(F(Y)) - \psi j(Y)\|_2^2 \quad (2)$$

The VGG19 loss network is denoted by ψj, where j represents the number of layers in the network. VGG19 is a pre-trained neural network model that extracts high-level features of images, gets their feature representation in the network, and calculates the squared Euclidean distance. We choose the number of layers conv1-5 and the weights are assigned as [1/32,1/16,1/8,1/4,1].

Generate the adversarial loss using the original loss function in VI-CycleGAN. The VI-CycleGAN loss is defined as follows:

$$L_{Cycle}(G,F) = \|F(G(X)) - X\|_1 + \|G(F(Y)) - Y\|_1 \quad (3)$$

Therefore, the objective of VI-CycleGAN can be formulated as follows, where D is the discriminator and γ controls the

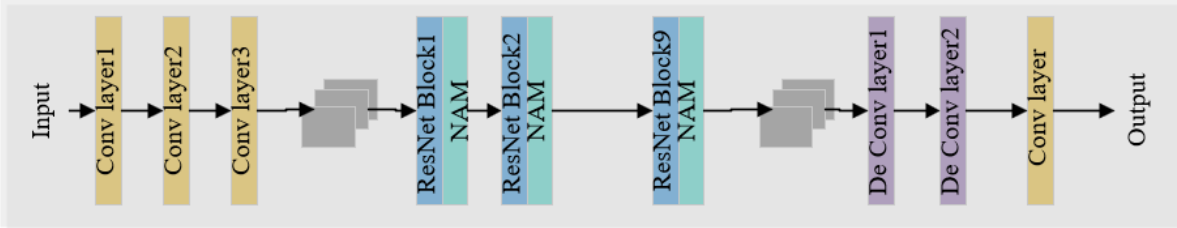

FIGURE 3. Generator network model.

weights of the loss function.

$$L(G, F, D_X, D_Y) = L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, X, Y) + \gamma_1 L_{Cycle} + \gamma_2 L_{Cycle-Edge} + \gamma_3 L_{Cycle-Perceptual} \quad (4)$$

IV. IMAGE DENOISE AND REGISTRATION METHODS

There is a certain degree of noise owing to the translation of visible images to fake infrared images. To improve the focus on pedestrians and reduce the influence of complex backgrounds. In this paper, we propose the use of guided filtering for IR images and fake IR images, and then use the robustness algorithms SURF and RANSAC to obtain accurate alignment results.

A. GUIDED FILTERING

Guided filtering [31] is an image filtering technique that processes the initial image using a guidance image. Although the fake infrared images generated by the VI-CycleGAN network can preserve edge features and details well, there are significant differences in detail and contour information between the fake infrared and infrared images. To address this issue, this study used fake infrared images as the guidance image and infrared images as the filtering input image, to enhance the common features in the registered image pair. Guided filtering also enhances the contrast and details of the image, making fake infrared images clearer and more realistic. The experimental results demonstrate that guided filtering can effectively preprocess images. The guidance image is denoted as I, the filtering input image is denoted as P, and Q represents the filtering result from P. We assume that the output image Q can be considered as a local linear transformation of the guidance image I.Meanwhile, we assume that the input image P is composed of Q combined with unwanted noise or texture. Defined by:

$$Q_i = a_k I_i + b_k, \forall i \in w_k \quad (5)$$

$$Q_i = P_i - n_i \quad (6)$$

where k is the midpoint of the local window, w_k denotes the linear window centered at pixel k, n_i is the noise or texture, and a_k and b_k are the linear coefficients, P represents the entire infrared image, while P_i represents the value of a specific pixel in the image. These pixel values are utilized to calculate the difference term at each pixel position in the objective function. The optimization problem is formulated to minimize the discrepancy between the output value of the fitted function Q_i and the filtered input image P_i, while simultaneously preserving the local linear model and better retaining the distinctive features of the infrared image. Defined by:

$$E(a_k, b_k) = \sum_{i \in w_k} ((a_k I_i + b_k - P_i)^2 + \varepsilon a_k^2) \quad (7)$$

where a_k and b_k are linear coefficients, ε is a constant, εa_k^2 is a regularization term that is used to control the smoothness. The concrete process of the proposed guided filtering method can be summarized as pseudo-code in **Algorithm 1**.

Algorithm 1 Guided Filtering Method

Input: guidance image I, filtering input image P
Output: filtering output Q
1 Initialize $a_k = 0, b_k = 0$;
2 for each pixel i:
Calculate the local mean: $m_I = Mean(w_i)$, $m_P = Mean(w_i)$;
Calculate the covariance: $c_I P = Covariance(w_i, I, P)$;
Calculate the local variance of I, $\sigma_I^2 = Variance(w_i)$;
Update the coefficients a_k and b_k:
$a_k = (c_I P - m_I m_P) / (\sigma_I^2 + \varepsilon)$;
$b_k = m_P - a_k m_I$;
3 return the filtering output Q;

B. ROUGH MATCHING

SURF [32] is an efficient and robust feature extraction algorithm. First, it detects points of interest by checking whether the determinant of the Hessian matrix is a local extremum and selects the optimal feature points. It then calculates the dominant orientation of each feature point and generates feature descriptors accordingly. Finally, the matching degree is determined by computing the Euclidean distance between the feature vectors of the two points, which yields an initial registration result.

C. FINE MATCHING

The RANSAC algorithm [33] is a random sampling consensus algorithm that iteratively trains the optimal parameter

model from a set of data containing mismatched points and eliminates incorrect matching points. The SURF algorithm is used to obtain the initial coarse registration points. However, there are many mismatches. Finally, the RANSAC algorithm is employed to remove the mismatched points, obtain the coordinates of the correct registration points, calculate the optimal single mapping transformation matrix between the 2D point pairs, and obtain the transformation parameters.

V. EXPERIMENTAL SETUP AND RESULTS

The experiment employed a server equipped with an NVIDIA GeForce RTX 2080 Ti GPU with 11GB of memory, and 256GB RAM to train the proposed network model. Training was performed on two datasets with distinct differences in image intensity distribution: RoadScene and MSRS infrared and visible image sets. The MSRS dataset was trained for 300 epochs, whereas the RoadScene dataset was trained for 200 epochs, with a batch size of 100 images, and an initial learning rate of 0.0001, using the Adam optimizer.

A. VI-CYCLEGAN QUALITATIVE ANALYSIS RESULTS

To verify the effectiveness of the VI-CycleGAN network model, we conducted a comparison experiment using the results of the CycleGAN network. The comparison results of converting visible images to fake infrared images are shown in **Figure 4.** The first column shows the visible images in the MSRS dataset. The second column shows the infrared image. The third column is the fake infrared image after the CycleGAN network translation. The fourth column shows the translation results of the proposed VI-CycleGAN. Globally, the translation results of the VI-CycleGAN network highlight the pedestrian and vehicle targets. Observed locally, both retain the edge contour features of the buildings and vehicles. The translated pedestrian information was richer in the last two sets of examples.

B. VI-CYCLEGAN QUANTITATIVE ANALYSIS RESULTS

To evaluate the effectiveness of the fake infrared images generated by the VI-CycleGAN network, three metrics were introduced, namely the mean square error (MSE), the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM). Among them, MSE calculates the difference between the pixel true values and predicted values. A lower MSE value indicates a higher similarity between the converted fake infrared image and the infrared image. PSNR is used as an objective measure to evaluate the similarity between two sets of images. A higher PSNR value indicates higher image quality. The SSIM metric comprehensively evaluates the quality of generated images based on brightness, contrast, and overall structure. The closer the SSIM value is to 1, the better the visual effect of the image. **Table 1** presents the average values of different evaluation metrics for the VI-CycleGAN algorithm on the MSRS test dataset.

Based on these evaluation metrics, we can observe that VI-CYCLEGAN outperforms CYCLEGAN in terms of MSE, PSNR, and SSIM. The generated images by VI-CYCLEGAN exhibit smaller pixel differences compared to the original images, indicating higher quality, and they are more similar to the infrared images in terms of structure and brightness.

TABLE 1. VI-CycleGAN quantitative analysis results.

NETWORK MODEL	MSE	PSNR	SSIM
CYCLEGAN	1.8697	23.3154	0.5124
VI-CYCLEGAN	1.1681	25.6201	0.6826

C. QUALITATIVE ANALYSIS OF THE REGISTRATION RESULTS

An example of the MSRS dataset is shown in **Figure 5**, where the light intensity is low, and an example of the RoadScene dataset is shown in **Figure 6**, where there is sufficient light relative to the MSRS dataset.

To validate the efficacy of the proposed algorithm, we compared it with the classical GFEMR algorithm [15], multimodal registration algorithm RIFT [16], and LNIFT algorithm [17] on two test sets. Figure 7 shows the registration results based on the MSRS dataset, where columns (a), (b), (c) and (d) correspond to the registration outcomes of the GFEMR, RIFT, LNIFT and proposed algorithms, respectively.

The results depicted in **Figure 7** provide clear evidence that the GFEMR algorithm (a) yields less favorable registration results due to the limited number of matching points, resulting in a relatively larger registration error. In contrast, the other algorithms exhibit comparatively better registration performance with a higher abundance of feature points, thereby achieving smaller errors. Notably, the LNIFT algorithm (c) demonstrates significant advancements over RIFT, as it noticeably increases the number of feature points and enhances the algorithm's robustness. The proposed algorithm(d) in this study surpasses alternative approaches by detecting features that are not identifiable in visible light images, thereby significantly bolstering its stability and robustness. **Figure 8** shows the registration results based on RoadScene dataset, where columns (a), (b), (c) and (d) correspond to the registration outcomes of the GFEMR, RIFT, LNIFT and proposed algorithms, respectively.

Figure 8 illustrates that the registration results of this dataset outperform those of the MSRS dataset, primarily due to its higher intensity, stronger similarity between the edge contours of infrared and visible light images, and lower noise levels. As the building outlines in this dataset are more distinct, the GFEMR algorithm (a) is capable of identifying accurate features, although the number of feature points is relatively limited. Both RIFT (b) and LNIFT (c) achieve a higher number of correctly registered points, but these types of algorithms struggle to accurately capture features with

FIGURE 4. Examples of experiments comparing VI-CycleGAN and CycleGAN.

FIGURE 5. Example of MSRS dataset.

significant brightness and representational differences. In contrast, the proposed algorithm(d) obtains a more abundant set of feature points and exhibits better robustness. These observations further demonstrate the effectiveness and reliability of the proposed algorithm.

D. QUANTITATIVE ANALYSIS OF THE REGISTRATION RESULTS

To verify the effectiveness of the proposed method, a quantitative evaluation was performed using precision, root mean square error (RMSE), as evaluation metrics.

FIGURE 6. Example of RoadScene dataset.

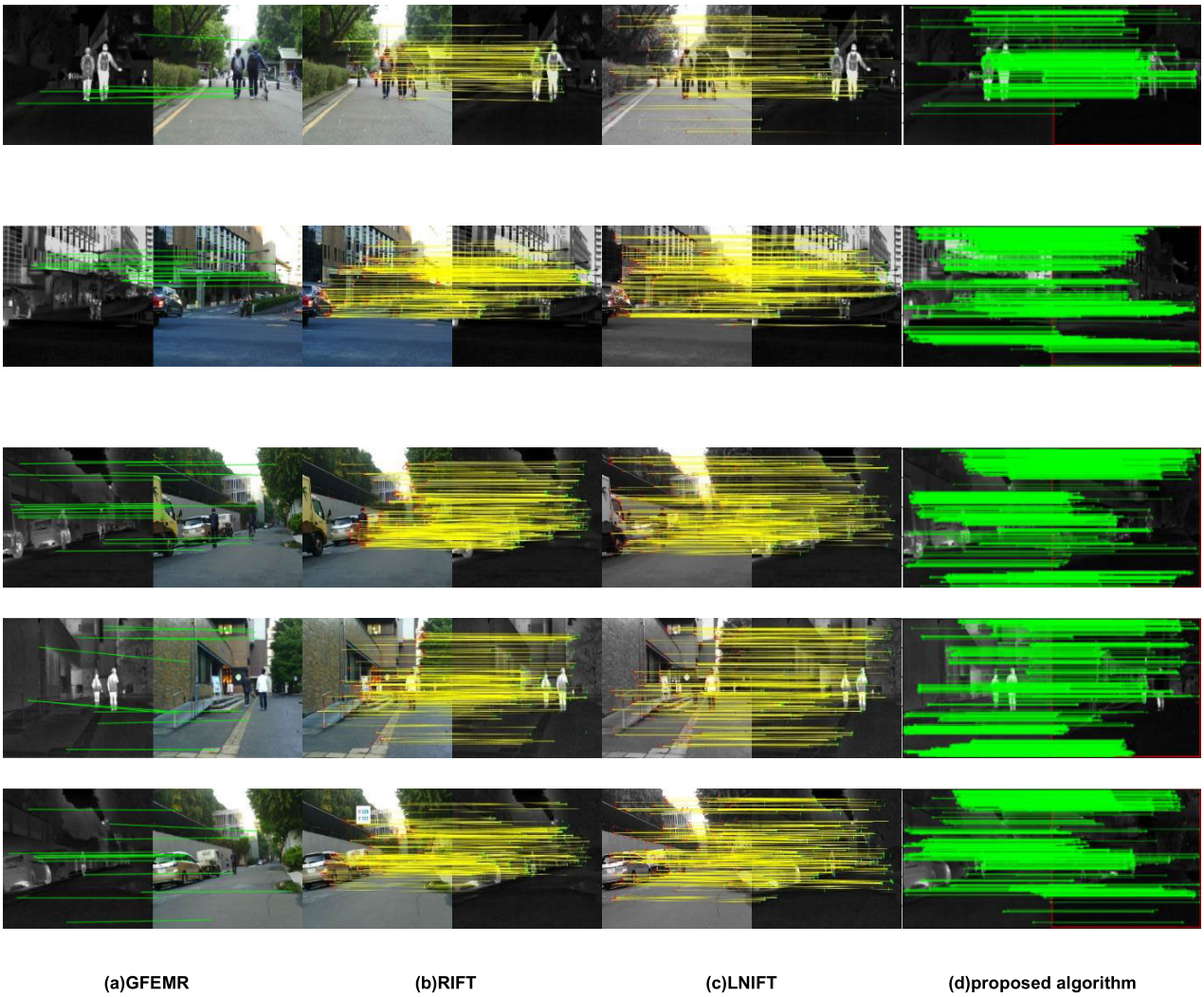

FIGURE 7. Example results from registering the MSRS dataset using different algorithms.

Precision is defined as the ratio of the number of correctly matched points to the total number of matched points, both of which are calculated from the registration results. The registration points that satisfy the following condition are

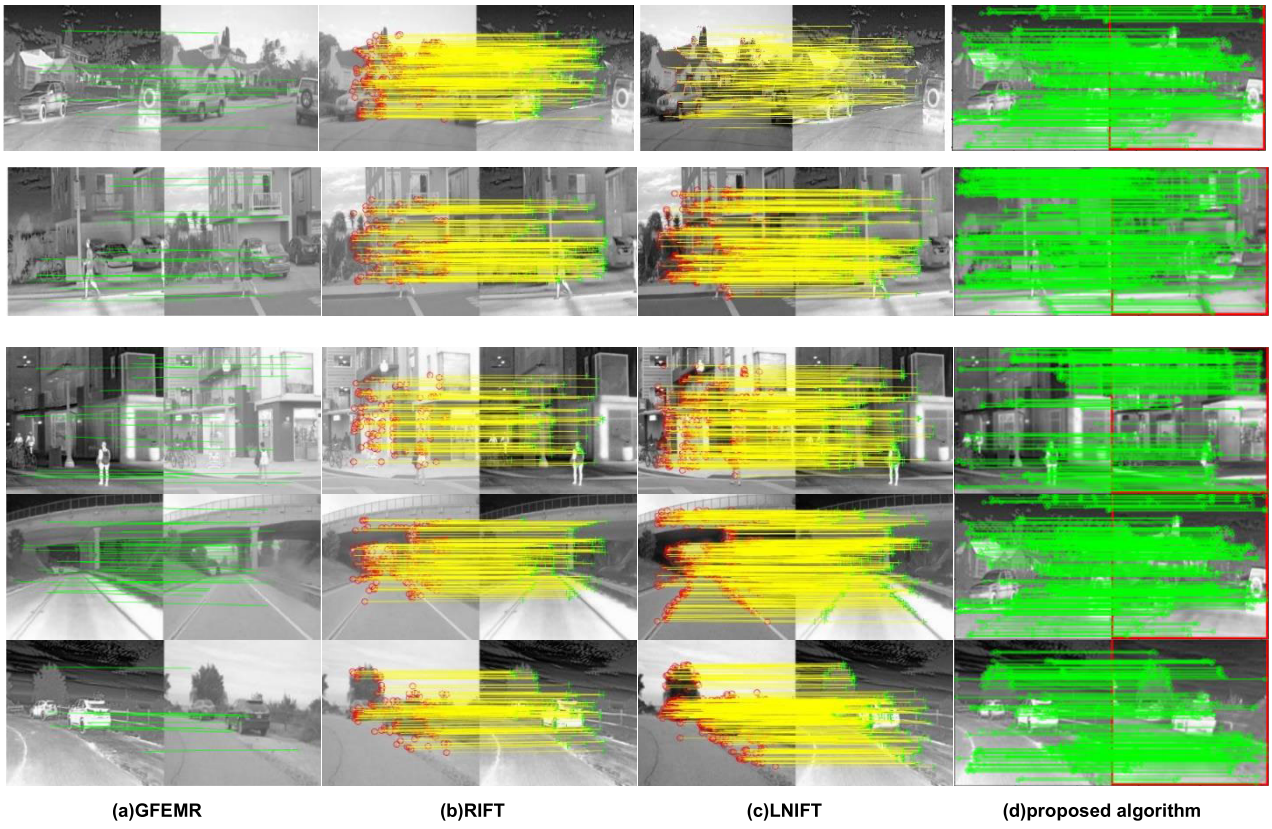

FIGURE 8. Example results from registering the RoadScene dataset using different algorithms.

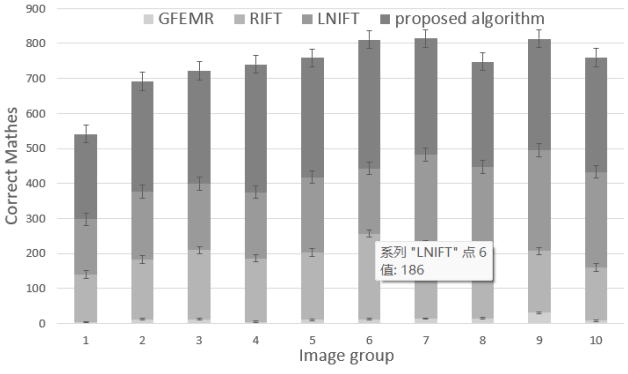

FIGURE 9. The number of correctly matched points for different algorithms.

considered Correct Matches:

$$\left\|(x_i, y_i) - \left(x_i', y_i'\right)\right\|_2 \leq 2 \tag{8}$$

The expression for registration precision is defined by:

$$Precision = \frac{CorrectMatches}{CorrectMatches + FalseMatches} \tag{9}$$

By using empirical methods to establish a threshold, the higher the registration precision, the more accurate the feature matching method.

The RMSE represents the root mean square error between the registered points obtained from the registration algorithm and the reference image matching points. The smaller the error, the better is the registration result. The expression for RMSE is defined by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} \left\|(x_i, y_i) - \left(x_i', y_i'\right)\right\|_2}{m}} \tag{10}$$

where m represents the final number of matched points, (x_i, y_i) are the coordinates in the reference image, and $\left(x_i', y_i'\right)$ are the transformed coordinates.

The registration precision of different algorithm test images was provided to quantitatively analyze the effectiveness of the proposed algorithm for image. The registration accuracies of the algorithms are listed in **Table 2**. **Table 2** demonstrates that the proposed algorithm outperforms the existing algorithms in terms of accuracy. In the test images, our algorithm consistently achieves an accuracy rate of over 85%. **Figure 9** shows the number of correctly matched points for the different algorithms.

Based on the comparison of correct matching points illustrated in **Figure 9**, it can be observed that the proposed algorithm achieves a significantly higher number of correct matching points. **Table 3** presents the RMSE values for the tested algorithms. It is noteworthy that a lower RMSE value corresponds to better registration performance.

TABLE 2. Registration accuracy of different algorithms.

Image group	Precision			
	GFEMR	RIFT	LNIFT	Proposed algorithm
1	25%	81.03%	85.87%	88.69%
2	48%	83.01%	88.18%	89.74%
3	48%	87.22%	87.91%	89.47%%
4	33.33%	88.73%	88.32%	90.57%
5	42.31%	84.58%	84.98%	87.88%
6	75%	90.41%	92.54%	95.58%
7	66.67%	91.01%	93.32%	96.51%
8	55.56%	90.20%	93.41%	97.10%
9	65.96%	88.89%	91.43%	93.53%
10	50%	91.57%	91.61%	95.61%

TABLE 3. Registration RMSE of different algorithms.

Image group	RMSE			
	GFEMR	RIFT	LNIFT	Proposed algorithm
1	5.8347	1.9889	1.3264	1.0881
2	6.1059	2.4125	2.2416	1.9145
3	5.2567	1.9563	1.3562	1.0321
4	6.5607	2.5462	1.9871	1.8548
5	6.4971	2.0156	1.8523	1.5245
6	5.1876	1.9785	1.5642	1.1210
7	4.2912	1.5623	1.2874	0.8522
8	4.4266	2.1452	1.5264	1.1465
9	3.3701	1.0564	1.0412	0.5632
10	4.8284	1.3546	1.1065	0.7936

Based on our analysis, registration points with error values less than 2 were considered correctly registered, while the remaining registration points were classified as misaligned. The results indicate that the GFEMR algorithm has the highest error values compared to the other evaluated methods. In contrast, the RIFT and LNIFT algorithms exhibit error values ranging from 1 to 3, and the proposed algorithm consistently achieves error values below 2. From our analysis, it can be concluded that our proposed algorithm is suitable for images with complex backgrounds and significant differences in image modalities.

VI. CONCLUSION

In this study, we propose a method based on modal conversion to align infrared images with visible images. First, because of the large representational differences between visible and infrared images, this study used the VI-CycleGAN network to convert visible images to approximate infrared images. Simultaneously, the NAM attention mechanism was added to the network to better capture the global information in the image and retain the image details. In this paper, we consider higher level semantic information and introduce hybrid loss to define the loss function, which can better preserve the content features of the original image. Guided filtering was used for image enhancement and noise processing. Finally, alignment results were obtained using the SURF and RANSAC algorithms. This study conducted experiments on two datasets and used evaluation metrics commonly used in academia for analysis. In the MSRS dataset, the average accuracy rate of the proposed method is about 85%. For the RoadScene dataset, the average accuracy rate of the proposed method was approximately 93%. The method in this study applies not only to visible and infrared images but also to other multimodal images with a certain degree of generalization.

The disadvantage of this method is that there is a certain degree of information loss during the modal transformation. If the infrared images in the dataset have fewer features, blurred edges, or no significant edges, the number of feature points may be smaller as pseudo-infrared images are generated, thereby affecting the alignment accuracy of the method. Although the VI-CycleGAN model can reduce the impact of this weakness, it remains a common problem in multimodal alignment. Therefore, future research should focus on two aspects:1) optimizing the generation effect of the VI-CycleGAN network to make it more generalizable and 2) improving the feature matching method to make the network focus on multi-scale features at different stages.

REFERENCES

[1] S. Saleem and A. Bais, "Visible spectrum and infra-red image matching: A new method," *Appl. Sci.*, vol. 10, no. 3, p. 1162, Feb. 2020.
[2] R. Feng, H. Shen, J. Bai, and X. Li, "Advances and opportunities in remote sensing image geometric registration: A systematic review of state-of-the-art approaches and future research directions," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 4, pp. 120–142, Dec. 2021.
[3] L. G. Brown, "A survey of image registration techniques," *ACM Comput. Surv.*, vol. 24, no. 4, pp. 325–376, Dec. 1992.
[4] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, no. 11, pp. 977–1000, Oct. 2003.
[5] D. G. Lowe, "Distinctive image features from scale-invariant key points," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Mar. 2004.
[6] Y. Li, X. Shi, L. Wei, J. Zou, and F. Chen, "Assigning main orientation to an EOH descriptor on multispectral images," *Sensors*, vol. 15, no. 7, pp. 15595–15610, Jul. 2015.
[7] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, "Asymmetric GAN for unpaired image-to-image translation," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5881–5896, Dec. 2019.
[8] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2868–2876.
[9] H. Emami, M. M. Aliabadi, M. Dong, and R. B. Chinnam, "SPA-GAN: Spatial attention GAN for image-to-image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 391–401, 2021.
[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[11] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image matching from handcrafted to deep features: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, Jan. 2021.

[12] Q. Li, G. Han, P. Liu, H. Yang, H. Luo, and J. Wu, "An infrared-visible image registration method based on the constrained point feature," *Sensors*, vol. 21, no. 4, p. 1188, Feb. 2021.
[13] S. Paul, U. K. Durgam, and U. C. Pati, "Multimodal optical image registration using modified SIFT," *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, vol. 1. Singapore: Springer, 2018, pp. 123–129.
[14] C. Gao and W. Li, "Multi-scale PIIFD for registration of multi-source remote sensing images," 2021, *arXiv:2104.12572*.
[15] J. Wang, J. Chen, H. Xu, S. Zhang, X. Mei, J. Huang, and J. Ma, "Gaussian field estimator with manifold regularization for retinal image registration," *Signal Process.*, vol. 157, pp. 225–235, Apr. 2019.
[16] J. Li, Q. Hu, and M. Ai, "RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform," *IEEE Trans. Image Process.*, vol. 29, pp. 3296–3310, 2020.
[17] J. Li, W. Xu, P. Shi, Y. Zhang, and Q. Hu, "LNIFT: Locally normalized image for rotation invariant multimodal feature matching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5621314.
[18] Q. Jiang, Y. Liu, Y. Yan, J. Deng, J. Fang, Z. Li, and X. Jiang, "A contour angle orientation for power equipment infrared and visible image registration," *IEEE Trans. Power Del.*, vol. 36, no. 4, pp. 2559–2569, Aug. 2021.
[19] Q. Wang, X. Gao, F. Wang, Z. Ji, and X. Hu, "Feature point matching method based on consistent edge structures for infrared and visible images," *Appl. Sci.*, vol. 10, no. 7, p. 2302, Mar. 2020.
[20] D. Zhao, "Rapid multimodal image registration based on the local edge histogram," *Math. Problems Eng.*, vol. 2021, pp. 1–9, Jun. 2021.
[21] P. A. Legg, P. L. Rosin, D. Marshall, and J. E. Morgan, "Feature neighbourhood mutual information for multi-modal image registration: An application to eye fundus imaging," *Pattern Recognit.*, vol. 48, no. 6, pp. 1937–1946, Jun. 2015.
[22] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, "Robust feature matching for remote sensing image registration via locally linear transforming," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015, doi: 10.1109/TGRS.2015.2441954.
[23] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 172–189.
[24] 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.
[25] Y. Liu, Z. Shao, Y. Teng, and N. Hoffmann, "NAM: Normalization-based attention module," 2021, *arXiv:2111.12419*.
[26] M. Aderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–12.
[27] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[28] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "ERN: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sens.*, vol. 10, no. 9, p. 1339, Aug. 2018.
[29] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[30] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," 2022, *arXiv:2205.11876*.
[31] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.
[32] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *Proc. Eur. Conf. Comput. Vis.*, vol. 110, Jul. 2006, pp. 404–417.
[33] Z. Hossein-Nejad and M. Nasri, "A-RANSAC: Adaptive random sample consensus method in multimodal retinal image registration," *Biomed. Signal Process. Control*, vol. 45, pp. 325–338, Aug. 2018.

YUAN WANG was born in Shaanxi, China, in 1998. She received the B.S. degree in software engineering from Weinan Normal University, in 2021. She is currently pursuing the M.S. degree with the School of Computer Science and Engineering, Xi'an University of Technology. Her research interests include computer vision and target recognition.

XIANGYANG LIANG received the B.S. degree from the Nanjing University of Science and Technology, Nanjing, China, in 1996, the M.S. degree in computer science and technology from Xi'an Technological University, Xi'an, China, in 2004, and the Ph.D. degree in computer simulation from Northwestern Polytechnical University, Xi'an, in 2008. Since June 1996, he has been with the School of Computer Science and Engineering, Xi'an Technological University. He is currently a Professor of computer science and technology. His current research interests include computer vision, big data analysis, artificial intelligence, system modeling, and distributed interaction simulation.

LEI CHEN received the M.S. degree in control theory and control engineering from Northeastern Forestry University, China, in 2006, and the Ph.D. degree in optical engineering from the Xi'an University of Technology, China, in 2022. He is currently a Lecturer with the School of Computer Science, Xi'an University of Technology. He has authored more than 40 research articles. His research interests include image processing, machine learning, and information fusion. He has received two awards for scientific progress from the Xi'an University of Technology.