International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2024): 6.623

Multimodal Feature Fusion for Wide-Angle Image Generation

Quanxing Peng

North China Electric Power University, School of Control and Computer Engineering, Beijing, 102206, China

Abstract: ***Image stitching combines visible light images from various perspectives to create wide-angle composites. However, adverse weather degrades these images, compromising stitching quality. Infrared sensors, which capture thermal radiation, excel in such conditions by highlighting targets. To overcome these challenges, we propose a multimodal fusion approach that integrates the robustness of infrared imaging with the rich textures of visible light. Our method uses a coarse-to-fine offset estimation based on infrared structural features and visible texture details, followed by a non-parametric Direct Linear Transformation for accurate geometric alignment, and finally fuses the stitched images to enhance scene perception. Tested on a real dataset of 530 multimodal pairs and a synthetic set of 200 pairs, our approach reduces average corner point error by 53%, eliminates ghosting, and boosts information entropy by 24.6% over DATFuse-UDIS++, demonstrating superior robustness and accuracy.***

Keywords: multimodal image fusion, infrared–visible stitching, wide-angle panorama, homography estimation, deep learning

1. Introduction

Due to the limited field of view, a single image cannot display the complete scene information. To address this issue, images from different viewpoints can be stitched together to obtain a composite image with a wider field of view. Image stitching is a fundamental step for further image understanding, and the quality of the stitching directly affects subsequent tasks, therefore an effective stitching algorithm is essential.

Traditional image stitching methods can be roughly divided into four steps: feature point detection, feature point matching, image registration, and image fusion. Among these steps, image registration is the key step influencing the stitching performance. Image registration estimates a 3×3 matrix to represent the deformation model from the target image domain to the reference image domain. However, in practical scenarios, targets at different viewpoints are often at different depth levels, and simply using a single global homography estimate for stitching often results in ghosting issues. To overcome this, existing algorithms 0, [2] perform position-adaptive transformations by learning, for example, an image can be divided into several small regions, each using a transformation model. This way, overlapping areas can achieve more effective registration results to some extent. Another class of traditional stitching methods is seam-driven [3]. These methods minimize the error of the seam to eliminate the influence of artifacts.

In recent years, deep learning has shown excellent performance in various tasks in the field of computer vision, and deep learning-based methods have also emerged in image stitching tasks [4]-**Error! Reference source not found.**[6]. These methods typically use deep convolutional neural networks to directly perform homography estimation and are only suitable for image stitching tasks from specific viewpoints, as they require a relatively small baseline range between different viewpoint images, thus limiting their generalization ability.

Considering the shortcomings of both traditional image stitching algorithms and deep learning-based stitching algorithms, we propose a multimodal data-based stitching method. First, the infrared sensor captures scene images by sensing thermal radiation in the environment, unaffected by occlusion, lighting, and other factors, avoiding the susceptibility to interference of conventional visible light imaging methods. However, due to the nature of infrared imaging, it has a relatively poor ability to perceive texture details in the target. As is well known, texture details are important as feature information in environmental perception, and visible light images can provide rich texture information through normal light reflection. Therefore, considering the complementary nature of infrared and visible light data, we integrates the advantages of both modalities during registration, and we adopt a multi-scale feature-pyramid structure. Deformation parameters are estimated from coarse to fine across multimodal data, and non-parametric direct linear transformation is used to estimate the deformation matrix. To fully utilize the feature information of different modalities, after obtaining the deformation parameters for the target and reference images, we use a reconstruction module to simultaneously achieve multi-view scene stitching and multimodal data fusion. In reconstruction module, deep scene features are mined to provide contextual semantic information, and shallow features are used to improve the upsampling of insufficient information, thus achieving more accurate and reliable fusion stitching results. This paper aims to develop and evaluate a multimodal infrared–visible image-stitching framework that delivers artifact-free wide-angle panoramas under challenging environmental conditions.

The main contributions are as follows:

- We introduce a wide-angle image generation algorithm based on multimodal data that harnesses the strengths of both infrared and visible imagery, overcoming environmental challenges faced by conventional methods and enriching scene information for improved perception.
- Our approach utilizes a multi-scale feature pyramid to regress global correlation loss and compute the transformation matrix non-parametrically, while the reconstruction module effectively compensates for information loss through contextual semantic awareness.

- Qualitative and quantitative evaluations on both real and synthetic datasets demonstrate that our algorithm achieves more accurate stitching with greater robustness and generalizability compared to existing traditional and deep learning-based methods.

2. Related Work

Traditional Image Stitching Methods

Traditional stitching methods are inherently prone to ghosting issues, and many approaches have been proposed to overcome artifacts. Gao et al. [2] introduced a dual homography estimation method for background and foreground targets. We use homography during alignment to achieve nonlinear distortion. Once the images are geometrically stitched, they are further processed to reduce artifacts in the stitching seam regions. Lin et al. [7] proposed a smooth deformation field, which reduces the impact of disparity in the stitching process by deforming the overlapping regions and performing extrapolation. This approach also meets the tolerance requirements for moving objects in different time domains to some extent. Zaragoza et al. [8] studied projection estimation when data does not fully satisfy the basic assumptions of the projection model and proposed the APAP (As Projective As Possible) method. It allows local non-projective biases to account for issues that do not meet the imaging assumptions, achieving seamless stitching via a novel moving direct linear transformation estimation technique. However, the APAP algorithm assumes that distortion changes are minimal in adjacent areas. In practical applications, the depth of adjacent regions can vary significantly, leading to disparity artifacts near the image boundaries. Chuang et al. [9] proposed a new parameter distortion algorithm that combines projection and similarity transformations, performing reasonable extrapolation for non-overlapping image areas based on projection. It retains the perspective information of the original image while achieving good alignment accuracy. Lin et al. [10] considered the smoothness of the deformation field while ensuring local image transformations, reducing the curvature and artifacts in the stitched image, thus improving the overall stitching result. Lee and Sim [11] introduced a deformation residual vector to distinguish matched features on different depth planes. For images with large disparity, the algorithm achieves more precise alignment by using corresponding homography estimates to distort different planes.

At the same time, another category of traditional stitching methods focuses specifically on the stitching regions and has also achieved notable performance in recent years. In the work of Gao et al. [2], a seam loss based on homography was proposed to measure the discontinuity between the distorted target image and the reference image, and the homography with the minimum seam loss was selected to achieve optimal stitching. It estimates geometric transformations not based on the best fit of feature correspondences but evaluates the quality of the transformation based on the visual quality of the seam cut. Experimental results show that the new image stitching strategy usually produces better perceptual results than existing methods, especially for challenging scenes. Zhang and Liu [3] introduced content-preserving warp (CPW) to align the re-looked areas for small local adjustments, while using homography to maintain the global image structure. Specifically, the method employed a hybrid alignment model that combined homography and content-preserving warping. This not only provided more accurate disparity estimation but also avoided local distortion issues. Furthermore, the method developed a seam detection approach that estimates reasonable seams from roughly aligned images by considering geometric alignment and image content. The resulting homography is then used to pre-align the input images, followed by local refinement using content-preserving warping. The aligned images are finally combined using a standard seam-cutting algorithm and multi-band blending. Unlike pixel alignment in overlapping regions, Lin et al. [12] used the estimated seams to guide the optimization of local alignment processes, improving seam quality with each iteration. Additionally, a new structure-preserving deformation method was introduced to preserve significant curves and line structures during deformation. These strategies greatly enhanced the effectiveness of the method in handling various challenging images with large disparities. Jia et al. [13] explored global collinear structures and incorporated them into the objective function to guide the balancing of features required for image deformation, thus preserving both local and global structures while reducing distortion. Liao et al. [14] addressed low-quality pixels in seams by separating prominent image blocks in the aligned images and performing local alignment using modified dense correspondences extracted via SIFT flow.

Deep Learning-Based Image Stitching

With the superior performance of Convolutional Neural Networks (CNNs) in computer vision, many researchers have explored their application in image stitching. DeTone et al. [15]were pioneers in integrating homography estimation into a CNN framework, estimating an eight-degree-of-freedom homography matrix to warp images from different perspectives into a unified spatial domain. However, due to the inherent limitations of linear transformations, this approach is mainly suitable for small-baseline scenes with minor parallax and struggles when dealing with significant depth variations, leading to poor perceptual quality and seam continuity issues. To overcome these challenges, Nie et al. [4] developed a deep image stitching method capable of handling arbitrary viewpoints. Their framework comprises a deep homography module, a spatial transformation module, and a deep image refinement module that collectively work to minimize artifacts and seam discrepancies via a structure-to-content stitching strategy. In an effort to improve generalizability, Nie et al. [5] later incorporated a local thin-plate spline interpolation to allow for more flexible warping and applied an iterative strategy for adaptive deformation across different resolutions. Unfortunately, due to the difficulty of obtaining ground-truth stitching labels in real-world settings, a fully supervised dataset for image stitching has yet to be established.

3. Methods

3.1 Net structure

Based on the above analysis, conventional visible light image stitching is easily influenced by environmental factors, such as rain, snow, fog, dust, low light, strong light, occlusion, etc.

In contrast, infrared sensors, which capture thermal radiation for imaging, are less affected by environmental factors. However, infrared sensors can only capture large-scale structural information of the scene, lacking the ability to represent fine details within the scene. This limitation makes it challenging for conventional algorithms to perceive the scene effectively. On the other hand, visible light images, captured via reflected light, can present detailed information

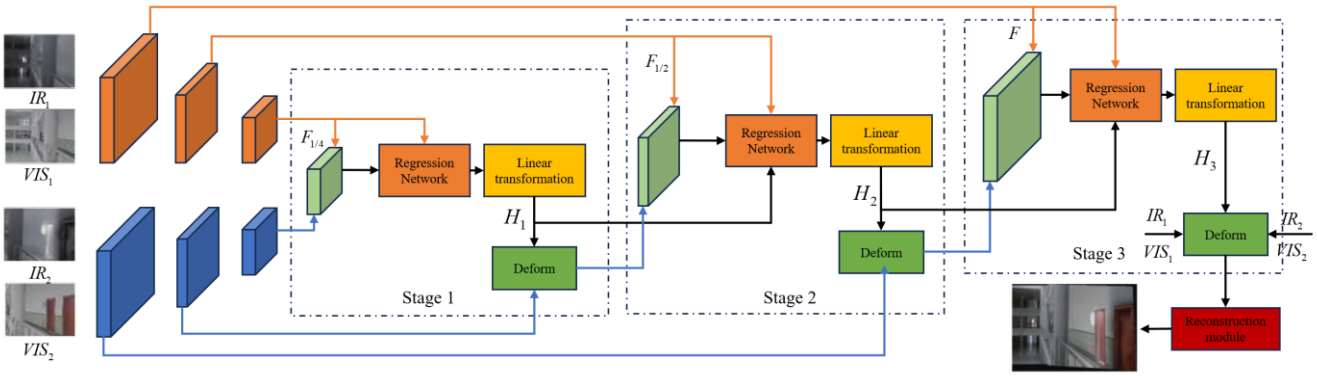

Figure 1: Workflow of the proposed method

well and complement the features of infrared imaging. Therefore, to generate wide-angle images under arbitrary scenes and perspectives, we propose a multi-modal data stitching method based on infrared and visible light images. Specifically, infrared image pairs (IR) and visible image pairs (VIS) captured from different viewpoints are first input into a multi-scale learnable pyramid network, which estimates the scene shift of specific modalities from coarse to fine. To achieve feature complementarity between the two modalities and highlight the advantages of multi-modal processing, the predicted latent offsets from the two modalities are fused to obtain a more accurate homography matrix. Here, the structural features of the infrared data can be combined with the detailed content of the visible light data. Then, the deformed images undergo global context-guided image reconstruction. Not only does it provide stitching results for scenes captured from different perspectives, but it also integrates information from multiple modalities, producing wide-angle images that are rich in information, clear, and accurate without artifacts. The overall network structure is shown in Figure 1. The entire network can be divided into three stages, which complement each other and achieve deformation parameter estimation from coarse to fine.

3.2 Multi-Scale Pyramid Based Homography Estimation

We first input images from different viewpoints of the same modality into the network and extracts features using two convolutional layers. Then, max pooling and two consecutive convolution operations are applied three times to construct a multi-scale feature pyramid, generating three feature sets F, $F_{1/2}$ and $F_{1/4}$, where each operation reduces the feature size by half.

The homography matrix estimation starts from the smallest-scale feature $F_{1/4}\,(16\times16)$. First, global correlation is computed at this scale to capture pixel-wise global feature relationships. A regression network, consisting of three convolutional layers and two fully connected layers, then predicts eight coordinate displacement parameters, which are used to compute the homography matrix in a non-parametric manner:

$$\Delta_1 = R\left(F_{A,1/4}, F_{B,1/4}\right) \tag{1}$$

$$H_1 = DLT\left\{R\left(F_{A,1/4}, F_{B,1/4}\right)\right\} \tag{2}$$

where $F_{A,1/4}, F_{B,1/4}$ are features from two viewpoints, $R(\cdot)$ represents the regression network, and $DLT\{\cdot\}$ denotes the non-parametric transformation calculation.

After computing the first-stage homography matrix H_1, it is applied to the second-stage feature $F_{B,1/2}$ for preliminary warping. Then, local correlation computation, regression, and transformation are performed again to obtain the second-stage homography matrix H_2, which provides a more refined transformation.

To further enhance performance, the method employs a three-stage recursive estimation. In the third stage, H_2 is used to transform $F_{B,1/4}$, followed by the same operations as in previous stages, ultimately obtaining the homography estimation for a single modality.

Since infrared and visible light images exhibit significant differences in content, the estimated homography matrices from different modalities are often inconsistent. To leverage their complementary advantages while mitigating individual limitations, we integrate the two homography matrices using a regression network with two fully connected layers, yielding a final homography estimation that fuses both modalities effectively

3.3 Wide-Angle Image Reconstruction Module

For the homography matrix obtained from the multi-scale pyramid structure, we deform the collected infrared and visible light image pairs to obtain pre-aligned image data suitable for stitching. However, the homography matrix performs best when both views lie on a common depth plane. In practical scenarios, it is difficult to meet such conditions. Additionally, because the two input images may be captured at different times and with different sensors, the brightness and color of the same spatial information often differ, leading to noticeable color differences in the seam area if the distorted images are directly integrated. Therefore, after the

deformation using the homography matrix, further reconstruction is required to generate a high-quality panoramic image with rich perceptual information while retaining the two perspectives.

Considering the differences in the information representation between infrared and visible light data, during the reconstruction stage, we simultaneously integrate both infrared and visible data to ensure that the final stitching result not only has a wider field of view but also contains complementary information from both modalities. To achieve this, we first introduce the U-net network, which extracts deep scene features to provide contextual semantic information and uses shallow features to further address the issue of insufficient upsampling information. Additionally, the attention mechanism is introduced in the skip connection process between the encoding and decoding layers to achieve

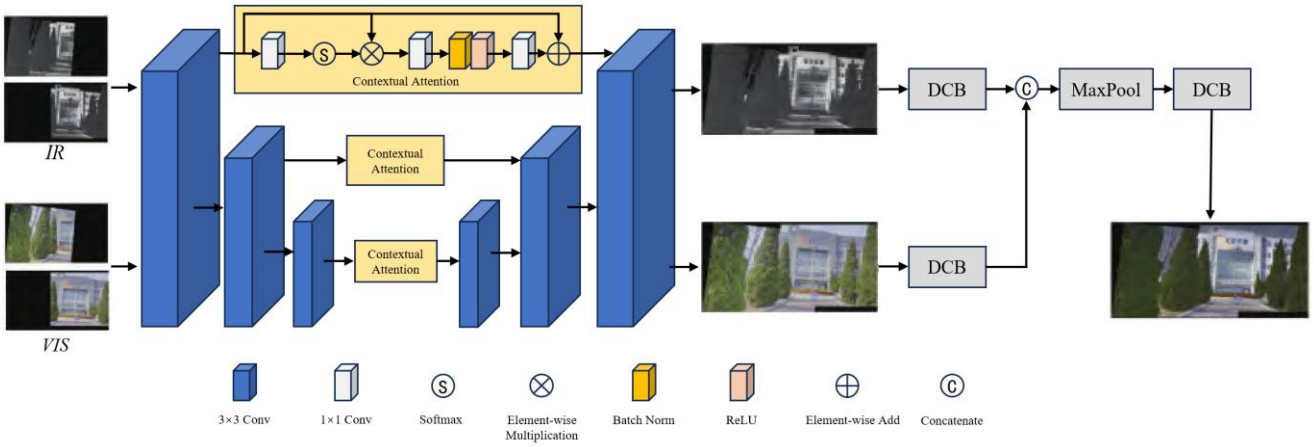

Figure 2: Detailed architecture of the image reconstruction module

accurate stitching in a self-supervised manner. The stitching results from both infrared and visible light data are then input into two dense connection blocks [16] for encoding. The encoded features from different modalities are fused through a weighted operation, and then decoded using another dense connection block to reconstruct the fused stitching result. A detailed network structure diagram is shown in Figure 2. In this way, by inputting images from different modalities and viewpoints, we generate a wide-angle image that combines the advantages of multi-modal data.

3.4 Loss Function

For the multi-scale homography estimation module, we adopt an unsupervised training approach. Specifically, a consistency constraint is applied to the common overlapping region of the deformed images from different viewpoints to ensure that the content in the overlapping region remains consistent after deformation and alignment. It indirectly ensures the accuracy of the deformation estimation. The loss function for this process is defined as follows:

$$\mathrm{L}_H = \lambda_1 \left|\varepsilon(I_A, I_B, H_1)\right|_1 + \lambda_2 \left|\varepsilon(I_A, I_B, H_2)\right|_1 + \lambda_3 \left|\varepsilon(I_A, I_B, H_3)\right|_1 \quad (3)$$

where I_A, I_B present two images from the same modality but different viewpoints. ε indicates the operation of selecting their overlapping region, can be formulated as:

$$\varepsilon(I_A, I_B, H_1) = I_A \times W(E, H_1) - W(I_B, H_1) \quad (4)$$

where $W(\cdot)$ represents the distortion operation that does not change the window size, and E is a mask of ones with the same size as I_B.

For the stitching part of the reconstruction module, a supervised approach is used. The stitching result is divided into two parts for reference: the stitching seam region and the non-stitching seam region. For the stitching seam region, an L1 norm constraint is applied. For the non-stitching seam region, to ensure that the features of the reconstructed image match as closely as possible with the features of the target image deformed by the homography matrix, we use a VGG (Visual Geometry Group) network for deep feature extraction. High-level features are then constrained to maintain perceptual consistency, thereby alleviating the stitching information mismatch caused by depth differences between the input images. Two loss functions are as follows:

$$\mathrm{L}_S = \left|I_A \times M_{S1} - I_S \times M_{S1}\right|_1 + \left|I_B \times M_{S2} - I_S \times M_{S2}\right|_1 \quad (5)$$

$$\mathrm{L}_C = \left|VGG(I_A) - VGG(I_S \times M_{C1})\right|_2 + \left|VGG(I_B) - VGG(I_S \times M_{C2})\right|_2 \quad (6)$$

4. Experiments

Experiments are implemented using TensorFlow on an RTX 3090 GPU. The input images are first passed through the multi-modal homography estimation network to obtain the homography matrix for initial alignment. The aligned images are then input into the reconstruction network for smooth seam transitions. For pretraining the homography network, the batch size is 4, with 50 epochs, 128×128 pixel input, and an initial learning rate of 1e-4. Loss function weight coefficients are set to λ_1=16, $\lambda_2 = 14$, and $\lambda_3 = 1$. For the feature reconstruction network, the batch size is 8, with 20 epochs, 128×256 pixel input, and an initial learning rate of 1e-3. Weight coefficients for L_S and L_C are set to 1 and 1e-5, respectively. After pretraining, both networks are jointly trained with a batch size of 1, 30 epochs, an initial learning rate of 2e-5, and weight coefficients for L_H, L_S, and L_C set to 1, 1, and 1e-5. Adam optimizer is used with a learning rate

decay factor of 0.96.

4.1 Quantitative comparison

Table 1: Quantitative Comparison Results

Method	EN	SF	AG	BRISQUE
VFIS	5.89	9.994	3.301	0.492
RSFI	5.93	10.029	2.667	0.489
UDIS++	5.61	9.669	3.057	0.491
Our	**6.99**	**11.186**	**3.373**	**0.495**

To objectively evaluate the performance of different stitching algorithms, we use four image quality assessment metrics on a real-world database: entropy (EN) (Roberts et al., 2008), spatial frequency (SF) [17], average gradient (AG) [18], and blind/referenceless image spatial quality evaluator (BRISQUE). All four metrics show a positive correlation with image quality.

The experimental results are shown in
Table 1. Our method achieves the highest values in EN, SF, and BRISQUE, while slightly lagging behind the SPW method in AG. Overall, compared to existing deep learning-based stitching algorithms, our method demonstrates clear advantages and robust performance.

4.2 Qualitative comparison

Figure 3 shows a comparison of the results between our algorithm and deep learning-based methods. In both examples, the VFIS, RSEI, and UDIS++ algorithms fail to effectively utilize the content information from the two-view scenes, leading to noticeable registration issues in multi-view scenarios and a significant loss of information in the final stitching result. In contrast, our method produces better stitching outcomes, with smooth transitions in the overlapping regions and a clear, accurate reconstructed scene, providing strong support for subsequent algorithms in environmental perception.

Figure 3: Qualitative Comparative Experimental Results

5. Conclusion

In this paper, we propose a wide-angle image generation method based on multi-modal data fusion, which combines the structural details of infrared images and the texture features of visible light images, overcoming environmental limitations of traditional stitching algorithms. Using a multi-scale feature pyramid and non-parametric deformation matrix calculation, the method ensures precise and reliable stitching. The context-aware reconstruction module further compensates for missing information. Experiments on synthetic and real-world datasets show that the method provides stable performance, accurate stitching in overlapping areas, and realistic scene restoration, avoiding ghosting and deformation.

References

[1] Lou Z, Gevers T. Image alignment by piecewise planar region matching[J]. IEEE Transactions on Multimedia, 2014, 16(7): 2052-2061.
[2] Gao J, Kim S J, Brown M S. Constructing image panoramas using dual-homography warping[C]//CVPR 2011. IEEE, 2011: 49-56.
[3] Zhang F, Liu F. Parallax-tolerant image stitching[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 3262-3269.
[4] Nie L, Lin C, Liao K, et al. A view-free image stitching network based on global homography[J]. Journal of Visual Communication and Image Representation, 2020, 73: 102950.
[5] Nie L, Lin C, Liao K, et al. Parallax-tolerant unsupervised deep image stitching[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 7399-7408.
[6] Shen C, Ji X, Miao C. Real-time image stitching with convolutional neural networks[C]//2019 IEEE International Conference on Real-time Computing and Robotics (RCAR). IEEE, 2019: 192-197.
[7] Lin W Y, Liu S, Matsushita Y, et al. Smoothly varying affine stitching[C]//CVPR 2011. IEEE, 2011: 345-352.
[8] Zaragoza J, Chin T J, Brown M S, et al. As-projective-as-possible image stitching with moving DLT[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 2339-2346.
[9] Chang C H, Sato Y, Chuang Y Y. Shape-preserving half-projective warps for image stitching[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 3254-3261.
[10] Lin C C, Pankanti S U, Natesan Ramamurthy K, et al. Adaptive as-natural-as-possible image stitching[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1155-1163.
[11] Lee K Y, Sim J Y. Warping residual based image stitching for large parallax[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8198-8206.
[12] Lin K, Jiang N, Cheong L F, et al. Seagull: Seam-guided local alignment for parallax-tolerant image stitching[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016: 370-385.
[13] Jia Q, Li Z J, Fan X, et al. Leveraging line-point consistence to preserve structures for wide parallax image stitching[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 12186-12195.
[14] Liao T, Zhao C, Li L, et al. Seam-guided local alignment and stitching for large parallax images[J]. arXiv preprint arXiv:2311.18564, 2023.
[15] DeTone D, Malisiewicz T, Rabinovich A. Deep image homography estimation[J]. arXiv preprint arXiv:1606.03798, 2016.

[16] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.
[17] Eskicioglu A M, Fisher P S. Image quality measures and their performance[J]. IEEE Transactions on communications, 1995, 43(12): 2959-2965.
[18] Cui G, Feng H, Xu Z, et al. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition[J]. Optics Communications, 2015, 341: 199-209.