2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

MT-Net: A heterogeneous image matching method based on modality transformation

Min Nuo[1], Fan Wang[1], Xinrong Wu[1], Xueqi Cheng[1] and Xiaopeng Hu[1*]

[1] Dalian University of Technology, Liaoning Province, China
*Corresponding author. Email: huxp@dlut.edu.cn
Contributing authors: wangfan@dlut.edu.cn, wuxinrong@mail.dlut.edu.cn, chxqidlut@163.com

Abstract. This paper focuses on challenges in heterogeneous image matching. The matching accuracy of heterogeneous image pairs is lower than that of homologous image pairs. Existing methods have attempted adaptive improvements to address the challenges in heterogeneous image matching, but the accuracy still needs improvement. This is because heterogeneous image pairs exhibit significant differences, primarily due to their distinct imaging mechanisms. Regarding this issue, we propose an end-to-end hybrid framework that employs modality transformation for heterogeneous image matching. First, a modality transformation method based on style transfer is proposed to convert heterogeneous image pairs into pseudo-homologous image pairs. Second, we extract multiscale and multilevel discriminative features from the pseudo-homologous image pairs to enhance the repeatability and discrimination of keypoints. Third, a unified matching loss is proposed to optimize the method for generating pseudo-homologous images. This loss function improves the performance of the modality transformation module and even the entire network. The experiments indicate that the proposed MT-Net improves the mean match result by 0.9% - 3.5%.

Keywords: Heterogeneous Images, Image Matching, Style Transfer, End-to-end learning.

1 Introduction

Heterogeneous image matching focuses on matching images captured by different sensors or under different imaging conditions. These images have significant differences in resolution, spectral response, image quality, and other aspects. The differences in visual features make the matching and analysis work complex and challenging. The goal of heterogeneous image matching is to find commonalities among these differences to achieve effective association and information fusion between images. Heterogeneous image matching has a wide range of applications in visual navigation, target search, medical imaging [1], remote sensing [2], pattern recognition, and more. However, unlike more mature homologous image matching technology, heterogeneous image matching still faces many challenges due to the significant differences in heterogeneous images caused by different imaging mechanisms.

The challenges motivate us to consider alleviating the problems using a modality transformation method based on style transfer to generate pseudo-homologous image pairs. Style transfer uses neural networks to extract the content of one image and the style of another and then combines these two to obtain the final result. In terms of implementation, we input infrared and visible images into the modality transformation module to generate pseudo-infrared images, thereby converting the matching problem between infrared and visible images into a matching problem between infrared and pseudo-infrared images.

In this paper, we first implement the modality transformation between infrared and visible images using a Generative Adversarial Networks (GAN) [3] framework. Its purpose is to generate infrared and pseudo-infrared image pairs to reduce the matching difficulties caused by large image differences. Then, we extract multilevel and multiscale features from the infrared and pseudo-infrared image pairs to obtain keypoints with high repeatability and discrimination, achieving matching of pseudo-homologous images. During the training process, we also optimize the performance of the modality transformation using a unified loss to achieve better match results. The main contributions of this paper are summarized as follows:

- We propose an end-to-end hybrid framework employing a modality transformation for heterogeneous image matching. The modality transformation method based on style transfer enhances the matching accuracy by converting heterogeneous images into pseudo-homologous images.
- We propose a unified loss to further optimize the performance of the modality transformation and the entire network. This loss function includes a transfer loss and a matching loss.
- We show that MT-Net increases the matching performance by 0.9%–3.5%. Additionally, the unified loss we proposed has improved model performance by approximately 0.4%.

2 Related Work

2.1 Image Matching

Image matching methods include traditional approaches and learning-based approaches. Hand-crafted methods such as SIFT [4] and SURF [5] are classic traditional image matching methods. With the development and application of deep learning, Jahrer et al. [6] introduce two trainable models for the extraction of descriptors. Both are based on convolutional neural networks. MatchNet [7] adds a metric learning network to improve performance and reduce the size of the learned descriptors. L2-Net [8] addresses the problem of negative samples being several orders of magnitude more numerous than positive samples in patch matching. Building on this, HardNet [9] uses a hard negative sampling scheme to ensure that the distance between the selected negative samples is minimized. LF-Net [10] based on a Siamese network processes the response maps to produce three dense maps, respectively representing the saliency, scale and orientation of keypoints. Based on the modifications to the structure of LF-Net, RF-Net [11] constructs receptive feature maps to achieve more effective keypoints detection. CS-Net [12] proposes a concurrent multiscale detector network, which consists of several parallel convolutional networks to extract multiscale and multilevel discriminative information for keypoint detection. However, these methods do not adapt well to heterogeneous images with significant differences. We add a modality transformation module base on style transfer to convert the challenging problem of matching heterogeneous images into a pseudo-homologous image matching problem.

2.2 Modality Transformation

The method of modality transformation we proposed is based on style transfer. MBS-GAN [13] proposes a novel Multi-Branch Semantic GAN designed to synthesize infrared images with different semantic information using multiple generators. Additionally, a classifier based on ResNet [14] is utilized to determine which generator's output is most appropriate for a specific input image. InfraGAN [15] utilizes a Conditional GAN [16] to generate conditional thermal images for a given visible image, and incorporates the Structural Similarity Index Measure (SSIM) [17] in the generator's loss function to ensure that the generated infrared images are structurally similar to the input visible images. However, when they are applied to heterogeneous image matching, they can be further optimized by the results of downstream tasks. Therefore, we propose a unified matching loss to optimize the method for generating pseudo-homologous images. This loss function improves the performance of the modality transformation module and even the entire network.

3 Methods

Our proposed MT-Net consists of two parts: modality transformation and image matching, as shown in Fig. 1. We use MT-Net to match visible and infrared images. First, we employ modality transformation based on a GAN framework to generate pseudo-homologous images for subsequent matching. Then, the image matching module extracts multiscale and multilevel discriminative features. We jointly train the modality transformation and image matching parts, using a unified loss to further optimize the results of the modality transformation.

3.1 Pseudo-homologous Image Generator

In the modality transformation stage, we use GAN [3] to generate pseudo-homologous images. GAN consists of two networks, generator and discriminator, which compete against each other. The generator tries to generate samples that resemble the original data distribution, whereas the discriminator tries to detect whether samples are real or generated. The discriminator is also a neural network, similar in structure to the generator. Its purpose is to distinguish between real data and fake data produced by the generator. The task becomes image-to-image transformation when the GAN architecture is conditioned on an input image.

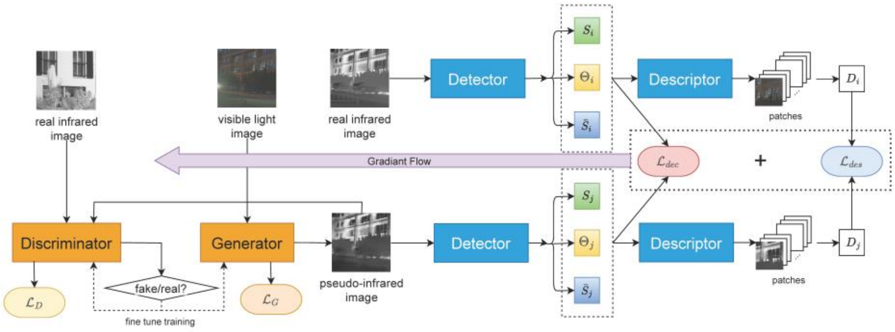

Fig. 1. Framework of our proposed end-to-end matching network with a modality transformation module. The modality transformation module with a discriminator and generator is used to generates pseudo-infrared images. Then, multiscale and multilevel discriminative features are extracted from the pseudo-homologous image pairs. During the training process, the modality transformation module is optimized not only with the losses from the generator (L_G) and discriminator (L_D) but also with the detector loss and descriptor loss (L_{det} and L_{des}) from the matching stage.

In our network, similar to InfraGAN [15], the generator improves the quality of infrared images by employing 2D convolutional layers with a stride of 2 for downscaling. It utilizes deconvolutional layers for upsampling without incorporating skip connections, relying on learned parameters to refine the image quality within each generator block. Additionally, since the input images are normalized to a range between -1 and 1, the generator concludes with a hyperbolic tangent activation function to produce pixel values that fall within the same input range.

Our discriminator uses bilinear interpolation for upsampling and maxpooling layers for downsampling, which helps reduce the number of learnable parameters and memory consumption. A fully connected layer with a single neuron is used as a normalization layer to constrain the output values between 0 and 1. Within the discriminator, two types of residual blocks are utilized: one for downsampling the input and the other for upsampling the input. Each of the residual blocks is equipped with 2D convolutional layers with 3×3 and 1×1 kernels, and they both apply the same size of padding. The upsampling block performs interpolation to increase the size of its input before applying the convolutional layers, while the downsampling block applies average pooling to decrease the size of the input after applying the convolutional layers. Finally, both blocks sum their two parallel paths on a per-channel basis and then pass the result along to the next block in the sequence.

We apply the GAN structure for modal transformation to match heterogeneous images. This method transforms the challenging problem of heterogeneous image matching into a problem of pseudo-homologous image matching. Our MT-Net further optimizes the effects of the modality transformation stage by using the loss from the image matching stage in a two-stage joint training process.

3.2 Keypoint Detector

The construction of scale-space response maps, denoted as $\{h^n\}$ where $1 \leq n \leq N$ and N is the total number of layers, is fundamental to keypoint detection. We use convolutional kernels of different sizes to extract multilevel feature maps. This approach endows the feature maps with discriminative capabilities at different hierarchical levels and provides them with receptive fields of different sizes.

To achieve this purpose, we employ N hierarchical convolutional layers to generate feature maps, each with several different sizes of receptive fields, with this range expanding as the convolution progresses. We then apply a 1×1 convolution to each feature map to produce multiscale response maps $\{h^n\}$. In our implementation, N is set to 10. Shortcut connections are added between each layer to facilitate network training without changing the receptive fields of the feature maps. In addition, we use a 1×1 kernel followed by instance normalization to generate the multiscale response maps, with all convolutions zero-padded to ensure the output size matches the input. This method uses both the abstract feature maps extracted from ResNet and the hierarchical structure to improve the scale space representation for keypoint detection.

In our approach, we identify keypoints based on high-response pixels from multiscale response maps $\{h^n\}$, constructing a keypoint score map accordingly. The keypoint detection process utilizes receptive feature maps to form these response maps. We employ two softmax operations to refine the score map S:

- The initial softmax enhances the response maps $\hat{h}^n$ by operating on a $15 \times 15 \times N$ window with zero padding.
- The subsequent softmax integrates these maps into the final score map S using the formula:

$$S = \sum_n \hat{h}^n \odot softmax_n(\hat{h}^n), \tag{1}$$

where $\odot$ denotes the Hadamard product.

For orientation estimation, we apply 1×1 convolutions to feature maps to generate orientation maps θ^n reflecting the sine and cosine of the orientations, and compute the angles using the arctan function. The final orientation map Θ is merged as

$$\Theta = \sum_n \theta^n \odot softmax_n(\hat{h}^n). \tag{2}$$

Similarly, the scale map $\bar{S}$ is derived by

$$\bar{S} = \sum_n \bar{s}^n \odot softmax_n(\hat{h}^n), \tag{3}$$

where $\bar{s}^n$ corresponds to the receptive field size of each feature map.

3.3 Descriptor Extraction

Our descriptor extraction module is designed with a series of convolutional layers. There are a total of seven layers involved. Each layer, except the last, is equipped with batch normalization to stabilize the learning process and ReLU activation functions to introduce nonlinearity. This helps the network learn complex patterns within the data. After passing through these layers, the output feature vectors are L2-normalized and have a dimension of 128. As for the loss function that guides the learning in this process, more details will be provided in the following sections. This structured approach allows us to effectively extract robust descriptors from the images.

3.4 Loss Function

We propose a unified loss to optimize our entire network which is calculated as follows:

$$L_{\text{total}} = \lambda_1 L_{\text{trans}} + \lambda_2 L_{\text{match}}, \tag{4}$$

where λ_1 and λ_2 are hyperparameters, and L_{trans} and L_{match} correspond to the loss of pseudo-homologous image generator and the loss of matching.

Transfer Loss. Our proposed transfer loss includes a generator loss and a discriminator loss.

Generator loss, denoted as L_G, is a composite of three terms: the standard conditional GAN loss (L_{cGAN}), an L1 loss (L_{L1}), and a loss based on the structural similarity index (L_{SSIM}). The generator's loss is formulated as follows:

$$L_G = L_{cGAN} + \lambda_3 \cdot L_{L1} + \lambda_4 \cdot L_{SSIM}, \tag{5}$$

where λ_3 and λ_4 are hyperparameters. The conditional GAN loss is given by

$$L_{cGAN} = -E_x \left[\sum_{i,j} \log\left(\left[D_{dec}\big(X, G(X)\big)\right]_{i,j}\right)\right] - E_x \left[\log\left(D_{enc}\big(X, G(X)\big)\right)\right]. \tag{6}$$

The SSIM loss is calculated as

$$L_{SSIM} = \frac{1}{m} \sum_{i=0}^{m-1} \big(1 - \text{SSIM}(G(X_i), Y_i)\big), \tag{7}$$

where m is the batch size, and SSIM measures the similarity between the generated image $G(X)$ and the ground truth image Y. Further details on the calculation of SSIM can be found in [17]. The generator's loss is computed assuming that the generated images are classified as real by the discriminator, thus driving the generator to produce more realistic images.

Discriminator Loss. The discriminator is trained with two loss terms. It captures the image-level discrimination loss while addressing the pixel-level discrimination loss. $L_{D_{enc}}$ represents the loss based on the image in the encoder output and $L_{D_{dec}}$ represents the loss based on pixels in the decoder output. The decoder output of the discriminator provides the probability that each pixel is real or fake. The loss function of the discriminator is expressed as

$$L_D = L_{D_{enc}} + L_{D_{dec}}, \tag{8}$$

where $L_{D_{enc}}$ and $L_{D_{dec}}$ are both based on cross-entropy loss, used for training the discriminator to distinguish between real and generated infrared images.

Matching Loss. In order to further optimize the modality transformation module for generating pseudo-infrared images based on the results of image matching, we propose a loss L_{match}:

$$L_{\text{match}} = L_{\text{det}} + L_{\text{des}}, \tag{9}$$

where L_{det} and L_{des} correspond to the detector loss and description loss. The impact of using L_{match} on the final matching results is discussed in Section 4.3.

Detector Loss. The primary objective of detector training is to ensure the repeatability of keypoint detection. The detector is trained once by

$$L_{\text{det}} = L_{\text{score}} + L_{\text{patch}}, \tag{10}$$

where L_{score} combines both RC-S-loss and RC-SD-loss followed by CS-Net [12] and L_{patch} is used to minimize the distance between matching descriptors. Specifically, we calculate the average distance between descriptor pairs of all keypoints as the loss function, a distance metric that is based on the similarity between descriptors.

Description Loss. The goal of descriptor learning is to align matching patches closely while separating nonmatching patches distinctly. Identical to CS-Net, we utilize effective triplet loss in conjunction with hard sample mining and neighbor masking for descriptor training. Triplet training samples are formed from both positive and negative samples. The triplet loss function is designed to ensure that there is a margin m by which the distance between the descriptors of negative samples exceeds the distance between the descriptors of positive samples:

$$L_{\text{des}} = \frac{1}{k}\sum_{i\in\text{keypoints}} \max\left(0, m + D(d_i, d_i') - D\left(d_i, d_j'\right)\right), \tag{11}$$

where k is the number of keypoints we have chosen, d_i' corresponds to the matching descriptor of d_i, and d_j' is a nonmatching descriptor. To address the limitations of randomly selected negative samples, which can lead to a slow training process and suboptimal performance, we select hard negative samples that are very close in distance.

4 Experiments

4.1 Datasets and Training

We verify the image matching performance on both VeDAI dataset [18] and RGB-NIR Scene dataset [19] to evaluate the effectiveness of our approach. VeDAI is a dataset for vehicle detection in aerial imagery. The images contained in the database exhibit different variability such as multiple orientations, lighting/shadowing changes, specularities or occlusions. RGB-NIR Scene dataset consists of RGB and near-infrared (NIR) images captured using visible and NIR filters. The most previous work uses 70% of the datasets for training, 20% for validation, and 10% for testing during the training phase. Within the 70% of the datasets used for training, we allocate 50% to generator training and the remaining 50% to the entire model training to prevent the generator from memorizing. The entire network is validated on the 20% of the datasets and tested on the 10% of the datasets.

For optimization, we use Adam [20] assigning a learning rate of 2×10^{-4} for the generator and 2×10^{-6} for the discriminator. We train the descriptor twice and the detector once, starting with an initial learning rate of 0.1 which decreased by 0.1 every five epochs. Batchsize is set to 8. We assign the values of λ_1 and λ_2 to 1 and the values of λ_3 and λ_4 to 100.

4.2 Evaluation Metric

This study utilizes the matching score (MS) [21] to assess image matching performance. The matching score is defined as $MS = N_r \, / \, N_m$ where N_r is the number of correct matches and N_m is the total number of matches, with the condition that $N_r \leq N_m$. A match is considered incorrect if the spatial distance between the predicted and ground truth points exceeds 5 pixels. The average number of correct matches, denoted as $\overline{N}_r$, is also used to evaluate performance, with higher values indicating better matching. The matching strategy has a significant effect on the results. We report MS and $\overline{N}_r$ for three strategies: nearest neighbor (NN), nearest neighbor distance threshold (NNT), and nearest neighbor distance ratio (NNR). In the NN strategy, a match is found by identifying the closest descriptor. NNT requires the closest descriptor to be within a distance threshold T. NNR matches if the ratio of the distances between the first nearest neighbor and d_i to the second nearest neighbor and d_i is below a threshold R_T. In this article, $T = 1$ and $R_T = 0.7$.

Table 1. Comparison of different methods on VeDAI and RGB-NIR Scene datasets.

Methods	VeDAI	RGB-NIR Scene	Average
SIFT[4]	0.492	0.462	0.477
SURF[5]	0.494	0.458	0.476
L2-Net[8]+SURF	0.635	0.583	0.609
L2-Net+ORB[22]	0.713	0.643	0.678
Hard-Net[9]+SURF	0.663	0.621	0.642
Hard-Net+ORB	0.632	0.604	0.618
LF-Net[10]	0.623	0.603	0.613
RF-Net[11]	0.804	0.763	0.784
CS-Net[12]	0.857	0.801	0.829
Ours	**0.892**	**0.810**	**0.851**

4.3 Results and Ablation Study

Experiments show that our proposed MT-Net achieves good results on the VeDAI and RGB-NIR Scene datasets, as shown in Fig. 2 and Fig. 3. To some extent, matching on the pseudo-homologous image pairs generated after modality transformation reduces the probability of false matches and increases the number of correct matches, compared to direct matching on heterogeneous image pairs.

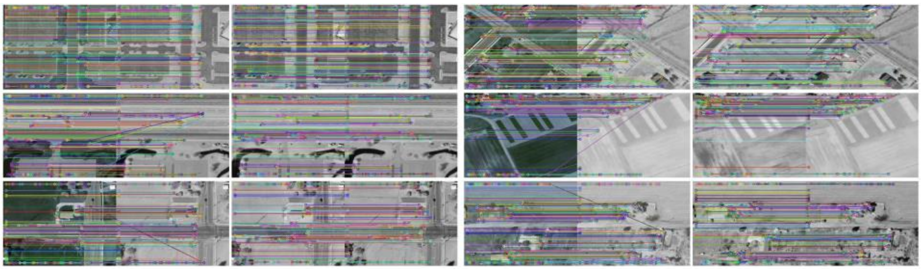

Fig. 2. Qualitative matching results on the VeDAI dataset, with correct matches.

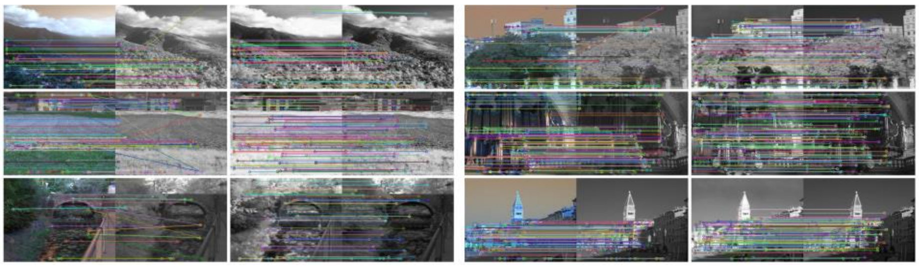

Fig. 3. Qualitative matching results on the RGB-NIR Scene dataset, with correct matches.

We compare MT-Net with other methods on different datasets. As shown in Table 1, our MT-Net achieve an average matching score that was higher than other methods on the VeDAI and RGB-NIR Scene datasets. This indicates that by transforming the heterogeneous image matching problem into a pseudo-homologous image matching problem through modality transformation, the effectiveness of heterogeneous image matching can be enhanced. Specifically, MT-Net uses the modality transformation method to convert input visible images into pseudo-infrared images. This approach reduces the challenges posed by large image discrepancies in heterogeneous image matching. By using the modality transformation method, MT-Net is able to take advantage of the more advanced techniques of homologous image matching, thereby enhancing the effectiveness of heterogeneous image matching.

Table 1 shows that our method performs better on the VeDAI dataset than on the RGB-NIR Scene dataset. Upon analysis, it is found that due to the more significant detail differences between near-infrared and visible images in the RGB-NIR Scene dataset, the quality of the generated pseudo-infrared images is not as good as that on the VeDAI dataset, as shown in Fig. 4. Therefore, the final matching performance on the

RGB-NIR Scene dataset is not as good as that on the VeDAI dataset. This further demonstrates that the results of heterogeneous image matching are influenced by the degree of image differences. Our approach, which focuses on reducing image differences, is therefore rational and effective.

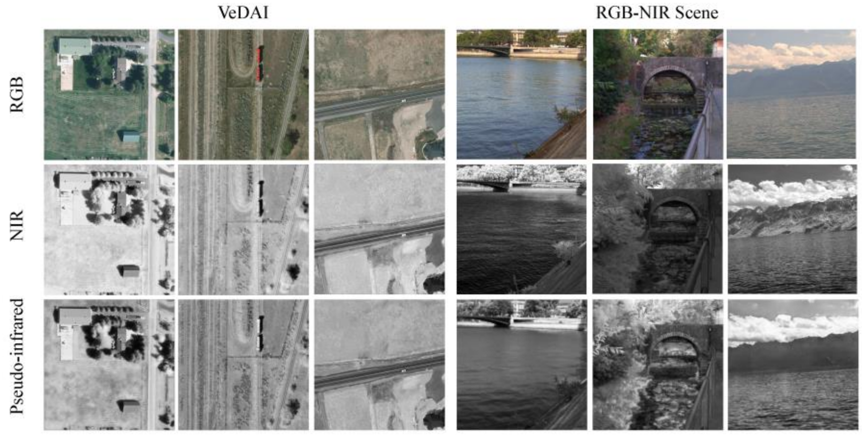

Fig. 4. The modality transformation effect on VeDAI dataset and RGB-NIR Scene dataset.

After analyzing the ablation study as shown in Table 2, we found that the MT-Net with added matching loss performed better on both datasets, which further indicates that the optimization of the modality transformation module has a positive impact on the matching results. The incorporation of the unified matching loss appears to have effectively guided the modality transformation process for subsequent matching tasks.

Table 2. Comparison of MT-Net with and without match loss on VeDAI and RGB-NIR Scene datasets.

Methods	VeDAI				RGB-NIR Scene			
	NN	NNT	NNR	Mean	NN	NNT	NNR	Mean
MT-Net(No L_{match})	0.715	0.954	0.996	0.888	0.583	0.892	0.976	0.817
MT-Net	0.731	0.947	0.998	**0.892**	0.596	0.886	0.981	**0.821**

5 Conclusion

We propose an end-to-end hybrid framework employing modality transformation based on style transfer for heterogeneous image matching. The modality transformation method addresses the problem of difficulties in matching heterogeneous images due to their significant differences by converting them into pseudo-homologous images. Moreover, we propose a unified loss to further optimize the effect of the modality transformation, enabling the modality transformation method to play a more effective role in the subsequent matching process. Our method has been proven effective through experiments, enhancing the performance of heterogeneous image matching. This success confirms the feasibility and effectiveness of applying the modality transformation method to such tasks. Therefore, we believe that the development of style transfer can add more methods and possibilities to the achievement of heterogeneous image matching.

References

1. Oh, S., Kim, S.: Deformable image registration in radiation therapy. Radiation oncology journal 35(2), 101 (2017)
2. Li, X., Yao, X., Fang, Y.: Building-a-nets: Robust building extraction from high- resolution remote sensing images with adversarial networks. IEEE Journal of Se- lected Topics in Applied Earth Observations and Remote Sensing 11(10), 3680– 3687 (2018)
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al.: Generative adversarial nets. Advances in neural information processing systems 27 (2014)
4. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Interna- tional journal of computer vision 60, 91–110 (2004)
5. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding 110(3), 346–359 (2008)
6. Jahrer, M., Grabner, M., Bischof, H.: Learned local descriptors for recognition and matching. In: Computer Vision Winter Workshop. vol. 2, pp. 103–118 (2008)
7. Han, X., Leung, T., Jia, Y., Sukthankar, R., Berg, A.C.: Matchnet: Unifying fea- ture and metric learning for patch-based matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3279–3286 (2015)
8. Tian, Y., Fan, B., Wu, F.: L2-net: Deep learning of discriminative patch descriptor in euclidean space. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 661–669 (2017)
9. Mishchuk, A., Mishkin, D., Radenovic, F., Matas, J.: Working hard to know your neighbor's margins: Local descriptor learning loss. Advances in neural information processing systems 30 (2017)
10. Ono, Y., Trulls, E., Fua, P., Yi, K.M.: Lf-net: Learning local features from images. Advances in neural information processing systems 31 (2018)
11. Shen, X., Wang, C., Li, X., Yu, Z., Li, J., Wen, C., et al.: Rf-net: An end-to-end image matching network based on receptive field. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8132–8140 (2019)
12. Quan, D., Wang, S., Huyan, N., Li, Y., Lei, R., Chanussot, J., et al.: A concurrent multiscale detector for end-to-end image matching. IEEE transactions on neural networks and learning systems 35(3), 3560–3574 (2022)
13. Li, L., Li, P., Yang, M., Gao, S.: Multi-branch semantic gan for infrared image generation from optical image. In: Intelligence Science and Big Data Engineering. Visual Data Engineering: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part I 9. pp. 484–494. Springer (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Özkanoğlu, M.A., Ozer, S.: Infragan: A gan architecture to transfer visible images to infrared domain. Pattern Recognition Letters 155, 69–76 (2022)
16. Mirza, M.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)

17. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full refer- ence image quality assessment algorithms. IEEE Transactions on image processing 15(11), 3440–3451 (2006)
18. Razakarivony, S., Jurie, F.: Vehicle detection in aerial imagery: A small target detection benchmark. Journal of Visual Communication and Image Representation 34, 187–203 (2016)
19. Brown, M., Süsstrunk, S.: Multi-spectral sift for scene category recognition. In: CVPR 2011. pp. 177–184. IEEE (2011)
20. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE transactions on pattern analysis and machine intelligence 27(10), 1615–1630 (2005)
22. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011)