Infrared-visible image patches matching via convolutional neural networks

Mao Yuanhong, Ma Zhong*, He Zhanzhuang

(Xi’an Microelectronics Technology Institute, Xi’an 710065, China)

Abstract: Infrared-visible image patches matching is widely used in many applications, such as vision-based navigation and target recognition. As infrared and visible sensors have different imaging principles, it is a challenge for the infrared-visible image patches matching. The deep learning has achieved state-of-the-art performance in patch-based image matching. However, it mainly focuses on visible image patches matching, which is rarely involved in the infrared-visible image patches. An infrared-visible image patch matching network (InViNet) based on convolutional neural networks (CNNs) was proposed. It consisted of two parts: feature extraction and feature matching. It focused more on images content themselves contrast, rather than imaging differences in infrared-visible images. In feature extraction, the contrastive loss and the triplet loss function could maximize the inter-class feature distance and reduce the intra-class distance. In this way, infrared-visible image features for matching were more distinguishable. Besides, the multi-scale spatial feature could provide region and shape information of infrared-visible images. The integration of low-level features and high-level features in InViNet could enhance the feature representation and facilitate subsequent image patches matching. With the improvements above, the accuracy of InViNet increased by 9.8%, compared with the state-of-the-art image matching networks.

Key words: infrared-visible image patches matching; convolutional neural networks; contrastive loss; triplet loss; multi-scale spatial feature integration

CLC number: TP391 **Document code:** A **DOI:** 10.3788/IRLA20200364

采用卷积神经网络的红外和可见光图像块匹配

毛远宏，马 钟*，贺占庄

(西安微电子技术研究所，陕西 西安 710065)

摘 要：红外和可见光图像块匹配在视觉导航和目标识别等任务中有着广泛的应用。由于红外和可见光传感器有不同的成像原理，红外和可见光图像块匹配更加具有挑战。深度学习在可见光领域图像的块匹配上取得了很好的性能，但是它们很少涉及到红外和可见光的图像块。文中提出了一种基于卷积神经网络的红外和可见光的图像块匹配网络。此网络由特征提取和特征匹配两部分组成。在特征提取过程中，使用对比和三重损失函数能够最大化不同类的图像块的特征距离，缩小同一类图像块的特征距离，使得网络能够更加关注于图像块的公共特征，而忽略红外和可见光成像之间差异。在红外和可见光图像中，不同尺度的空间特征能够提供更加丰富的区域和轮廓信息。红外和可见光图像块的

收稿日期:2020-09-20； **修订日期:**2020-11-04

基金项目:国家自然科学基金青年科学基金 (61905089); 中国航天科技集团公司第九研究院创新基金 (2019JY06); 中国航天科技集团钱学森青年创新基金 (2019 年度第 39 项)

作者简介:毛远宏，男，博士生，研究方向为计算机视觉、机器学习和人工智能。

通讯作者:马钟，男，高级工程师，博士后，研究方向为计算机视觉、机器学习和人工智能。

导师简介:贺占庄，男，研究员，博士生导师，博士，研究方向为嵌入式系统架构、计算机操作系统、计算机控制。

高层特征和底层特征融合可以有效地提升特征的表现能力。改进后的网络相比于先前卷积神经匹配网络,准确率提升了 9.8%。

关键词：红外和可见光图像块匹配；　卷积神经网络；　对比损失；　三重损失；　多尺度特征融合

0 Introduction

Infrared-visible image patches matching is a fundamental task of infrared-visible image processing. It compares the object or region by analyzing the similarity of content, features, structures, relationships, textures, and grayscales in infrared-visible images. The infrared-visible image matching is often used as a subroutine that plays an important role in a wide variety of applications, such as visual navigation[1−2] and target recognition[3−4].

Infrared-visible image patches is more challenging compared with traditional visible images. Since infrared and visible sensors use different imaging principles, the images taken by multiple sensors also have more differences than those by a single sensor. The edges of the object are blurred in infrared images. Less texture and color features are found in the object. The infrared-visible image pairs have significant grayscale distortion and illumination change.

Manual descriptors are used to extract features, such as SIFT[5], SURF[6], ORB[7], etc. The features extracted with the descriptors should have the invariance of illumination, rotation, scale, and affine. After feature extraction, image patches matching is predicted by comparison of features similarity. Most work is focused on improvements to infrared and visible image descriptors in the traditional infrared-visible image system. Sima[8] optimized the SIFT method for the infrared-visible image. Li[9] detected object edges and extracted the features of SURF to match the infrared-visible images. Chao Zhiguo[10] proposed a matching method based on histograms of oriented gradients used as the matching feature and the correlation coefficient used as a similar measure. Cao Zhiguo[11] adopted an approach to shape contexts for matching infrared-visible images based on their similar shape. Jiao Anbo[12] proposed the image matching algorithm using linear group geometric primitives for infrared and visible template matching.

Hand-craft descriptors need to improve continuously for new applications to extract efficient features. The feature extraction and similarity measure are two independent and unrelated stages, which cannot be optimized end-to-end. With the widespread application of deep learning in computer vision, the image patches matching based on deep learning has become a trend. MatchNet[13] extracts the image features from two CNN branches. It uses two full connection (FC) layers to determine whether the extracted features are similar. Deep Compare Network[14] compares the image patches by Siamese networks, 2-channels, and pseudo-Siamese models. Patch match networks[15] proposed improved architecture for two-channel and Siamese networks to compare the visible image patches. The networks above have achieved excellent performance in visible images. However, they do not solve the infrared-visible image patches matching well. The patches have different imaging principles. It is necessary to design a new deep neural network to achieve better performance in infrared-visible images matching.

This paper proposes an infrared-visible image deep matching network (InViNet) to tackle these challenges above. Two CNN branches extract the infrared and visible image features independently. The full connection layers compare their similarity.

In infrared-visible image patches matching, we think that the differences between unrelated patches are still more significant than those within similar patches, even if multi-sensors take the infrared and visible images. The feature extraction subnetwork uses the contrastive loss and triplet loss to maximize the distance of the feature between unrelated patches and minimize it within similar patches. It makes the distribution of the high-level feature more centralized within the intra-classes and more separate between the inter-classes.

For the infrared images, their regions and shapes still have essential references in the infrared-visible image matching. Integrating the spatial features with semantic features is necessary. We combine the multi-scale spatial features with the high-level features to enhance the performance. Compared to the previous CNNs, our method can increase the accuracy from 78.95% to 88.75%.

1 Infrared-visible image patches matching network

1.1 Overview of our network architecture

Our network mainly consists of two parts: the feature extraction network and the metric network, as shown in Fig.1. The feature extraction network is responsible for extracting features in infrared and visible images. The metric network mainly matches the feature’s similarities.

The feature extraction network extracts the distinguishing features of visible and infrared images. In the feature extraction network, infrared and visible images are input into two VGG16[16] branches, which constitute a Siamese network. To be compatible with the visible image ’s three channels, we copy an infrared image into three channels as another branch input. The weights are shared in the two branches. There are five blocks and two FC layers in a single VGG16 branch. A block consists of two or three convolution layers, an

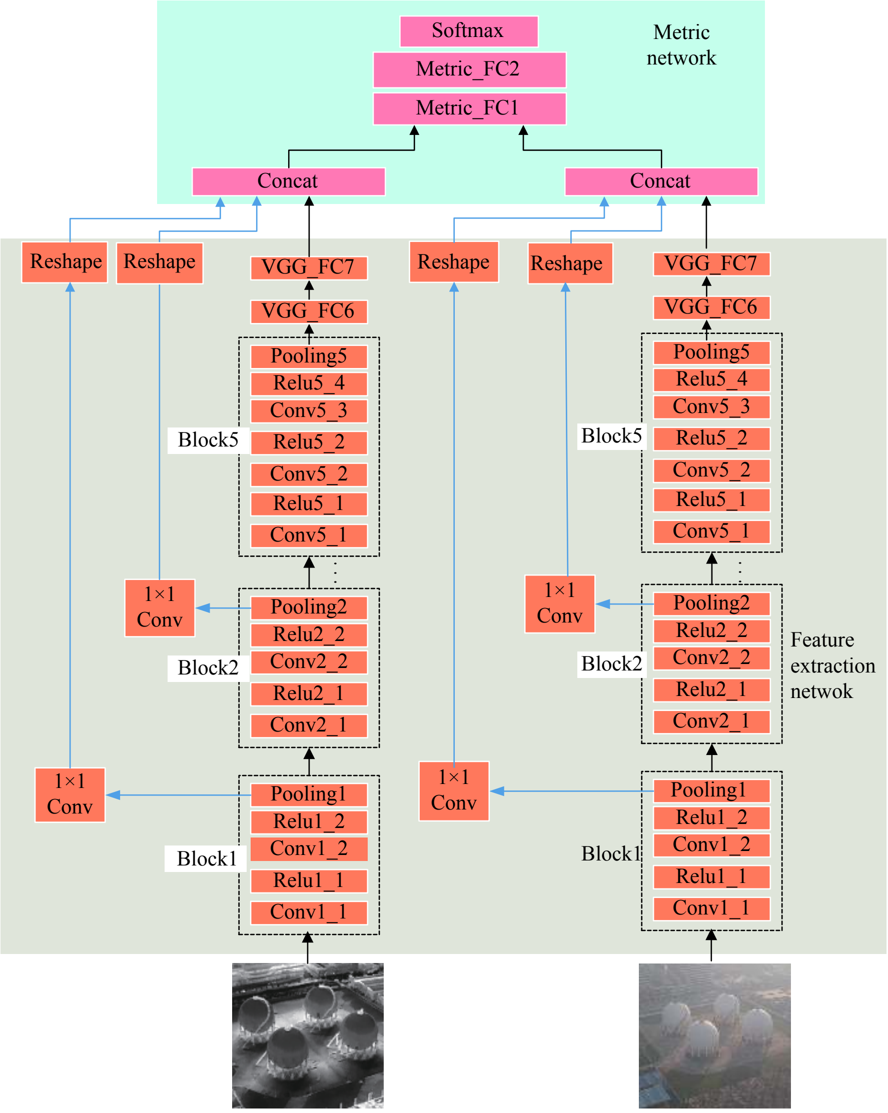


Fig.1 Infrared-visible image deep matching network. The black line with the arrow indicates the data-flow. The blue lines represent shortcut connections through the reshape layers. This figure describes the process of the infrared-visible image patches matching

activation layer, and a pooling layer. In a single VGG branch, we retain the original FC6 and FC7 layers in VGG16. There are two reasons for retaining the two FC layers. Firstly, the FC7 layer can produce a feature of 1×4 096 dimension, rather than 7×7×512 dimensions from the Conv5 block. It can significantly reduce the parameters and calculations in the metric network. Secondly, we find that the branch with FC layers has better performance than that without them in training.

For infrared and visible images, although their imaging principles are different, the same target is very similar in semantic features. Therefore, branches share network weights in network design. We believe that deep convolutional networks have strong feature representation capacity. It can extract common feature in infrared and visible images. Multiple network branches that traditionally use contrastive loss or triplet loss generally share weights. The shared weights can map high-level features to the same feature space for distance comparison.

The metric network is composed of two FC layers with softmax loss as the objective function. It estimates the probability of whether the visible image and the infrared image are similar or not. Ideally, if they match, the prediction is 1. If they don't match, the prediction is 0.

1.2 Multi-scale spatial feature integration

Compared with visible images, infrared images have no color and less texture information. The edges are usually blurred. However, the objects still have rough outlines and region information in infrared images. These outlines and shapes are common features in visible and infrared images. Therefore, we believe that their spatial information is essential in infrared images for image matching. It is necessary to integrate the spatial features with the semantic features to enhance feature representation.

On the other hand, it is feasible to propose features with multiple scales in the deep neural network's hierarchical framework. The features proposed from the low-level layers are similar to those extracted with the hand-craft descriptors, such as SIFT, SURF. As the CNN layers deepen, the feature maps less focus on the imaging difference. The semantic features gradually reveal in the high-level layers. In our network, the multi-scale features are input into the metric network. So, the metric network can use more comprehensive information to make similarity decisions. Each block in our network directly connects to the input of the metric network. It can preserve more multi-scale spatial information for similarity comparison in the metric network, as shown in Fig.2.

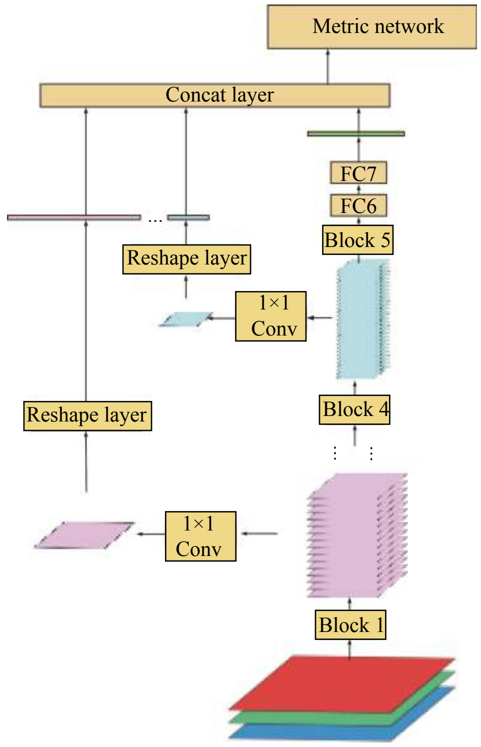


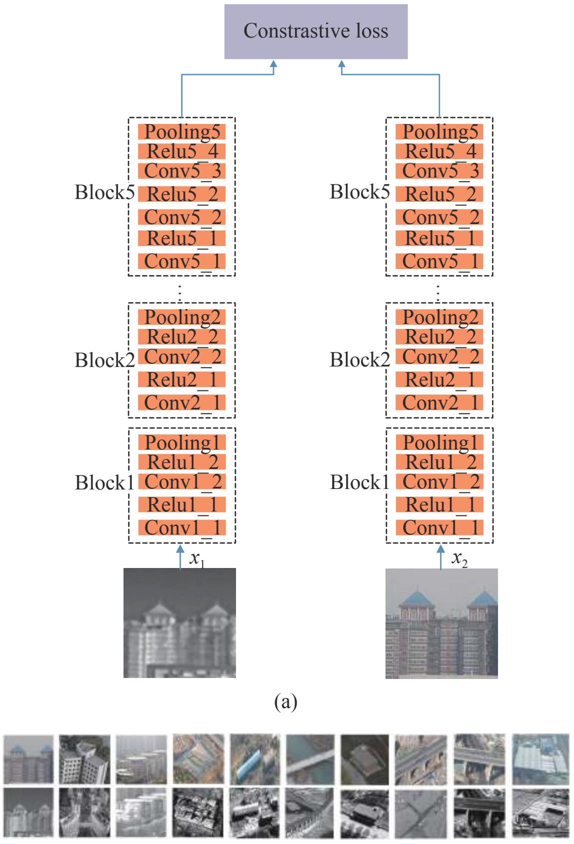
Fig.2 Multi-scale spatial feature integration in a single branch. The output feature map in each block shorts to the concatenation layer. The output of the concatenation layer is one input of the metric network

In multi-scale spatial feature extraction, two problems need to be solved. Firstly, the shortcut feature should maintain the original feature maps' size in each block to preserve spatial information. Secondly, the shortcut feature dimensions should not be too high after it reshapes into a vector. The great dimension eventually results in vast parameters and high computation in the metric network.

The 1×1 convolution is adopted in our network to solve the problems. The 1×1 convolution is widely used in GoogLeNet[17]. The multi-channel feature maps are compressed into a single-channel feature map, which preserves the spatial information and avoids the too high dimensions. To connect the features of different dimensions, they are converted to the vectors of length $N\times N$ with the reshape layer. N represents the size of the corresponding feature map. All multi-scale feature maps, including the semantic feature from the FC7, are concatenated as the input of the metric network. In the metric networks, its inputs include the infrared image branch and the visible image branch.

1.3 Two shared branches in feature extraction network

As shown in Fig.3, the network of feature extraction consists of two branches. Two branches are identical in structure and share weights. A visible image and an infrared image make up an image pair. The contrastive loss was first used for dimensionality reduction[18]. Here, the contrastive loss is used as the objective function to train the two branches.



(a)

(b)

Fig.3 (a) Feature extraction network architecture with the contrastive loss; (b) Input data for feature extraction network with the contrastive loss. The visual patches are in the first row. The infrared patches are in the second row. The positive samples are in odd columns. The negative ones are in even columns

The contrastive loss is shown in Eq.(1).

$$l(x_1,x_2)=\begin{cases} d(f(x_1),f(x_2)), & p_1=p_2\\ \max(0,\mathrm{margin}-d(f(x_1),f(x_2))), & p_1\neq p_2\end{cases} \tag{1}$$

where $d(f(x_1),f(x_2))$ represents the Euclidean distance of two sample features; p_1 is the label of input visual image; p_2 is the label of input infrared image. $p_1=p_2$ means a similar patches pair. $p_1\neq p_2$ means an unrelated patches pair. The margin is a threshold in Eq.(1). It represents the distance that should be separated from the unrelated features, at least. In our experiment, the margin is set to 1.

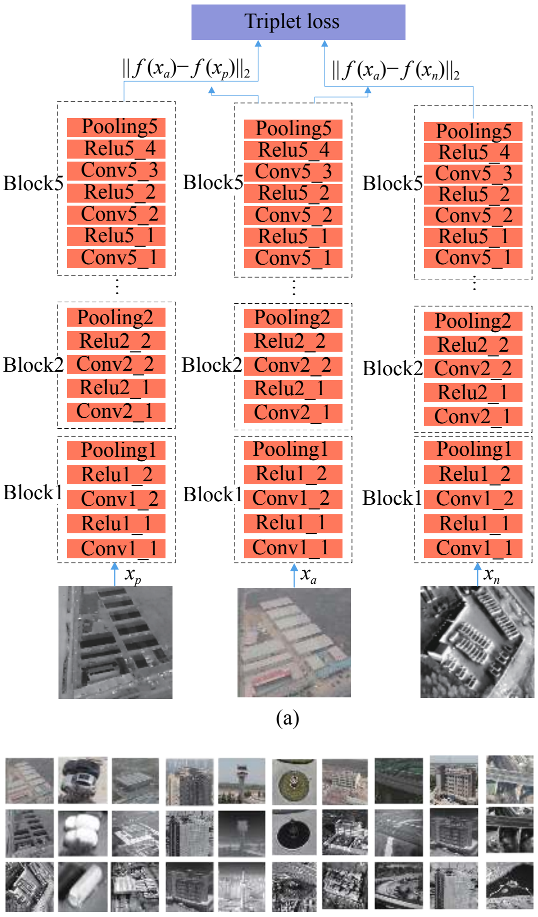
1.4 Triplet shared branches in the feature extraction network

As shown in Fig.4 (a), the network consists of triple branches. Three branches are identical in structure and share weights. A visible patch (anchoring sample), an infrared patch (positive sample), and another infrared patch (negative sample) form an image pair. We input a triple pair at a time to train the feature extraction network. The triplet loss was used for face recognition[19] firstly. Here, it is used as the objective function to train the triple branches.

The triplet loss shows in Eq.(2).

$$\max(d(f(x_a),f(x_p))-d(f(x_a),f(x_n))+\mathrm{margin},0) \tag{2}$$

The input data include anchoring sample (x_a), positive sample (x_p) and negative sample (x_n). $d(f(x_a),f(x_p))$ represents the Euclidean distance of the anchoring sample and the positive sample. $d(f(x_a),f(x_n))$ represents the Euclidean distance of the anchoring sample and the negative sample. By optimizing the function, the distance between the anchoring example and the positive example is less than the distance between the anchoring example and the negative example. The anchoring example is randomly selected from the sample set. The positive example and the anchoring example go to the same class,



(a)

(b)

Fig.4 (a) Feature extraction network architecture with the triplet loss; (b) Input data for feature extraction network with the triplet loss. The anchor patches are in the first row. The positive patches are in the second row. The negative patches are in the third row. Each column is triple patches input

while the negative example and the anchoring example belong to different classes.

$$d(f(x_a),f(x_p))+\text{margin}\leqslant d(f(x_a),f(x_n)) \qquad (3)$$

Eq.(3) illustrates that there is a margin between $d(f(x_a),f(x_p))$ and $d(f(x_a),f(x_n))$ to distinguish positive and negative samples. Unlike contrastive loss, the triplet loss function compares the distance between positive and negative samples in a forward and backpropagation process. Compared to visible image matching, the same object's imaging difference is also relatively large in multi-source image patches matching. So, it is found that a larger margin can achieve better performance in our experiment. The margin is set 3 to achieve the best performance.

2 Experiment

2.1 Data set

There are no available infrared-visible image patches matching datasets on the Internet, so we have to collect image pairs ourselves. In data acquisition, the visible camera is the default equipment in the DJI UAV. The infrared camera is manufactured by FLIR company. The wavelength of the infrared camera ranges from 7.5 to 13.5 μm. In terms of image resolution, the UAV acquires infrared and visible images at different altitudes. In the original image, the proportion of the same target to the image size is 0.8×, 0.5×, and 0.25×, respectively. In the following data preprocess, we crop the target area from the original images. The input images of the neural network resize to 224×224. Therefore, we use different resolution images during training and testing.

Our data set contains 2 000 images, falling into 25 classes. For scene selection, the target taken by UAVs should be different in shape and outline. The classes cover bridges, buildings, roads, parking lots, factories, houses, towers, gas storage tanks, etc., as shown in Fig.5. In the data set, the ratio of visible and infrared images is 1 ∶ 1. 80% of the images are used as training data. The rest images are used as test data. A sample includes an infrared patch and a visible patch. If the image pairs are similar, they are positive samples. Their ground truth is 1. If they are not similar, they are negative samples, and the ground truth is 0. In the training and test data set, the ratio of positive and negative samples is 1 ∶ 1.

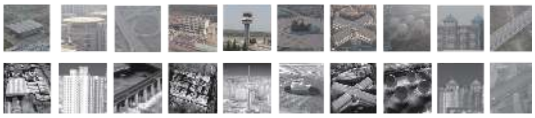


Fig.5 Infrared-visible image samples. Ten image pairs randomly was selected. The ground truth of the first five images is 0. The ground truth of the last five columns is 1

2.2 Experiment method

InViNet using two-stage training is better than the traditional classification network. In two-stage training, the feature network can improve the features representation. It can significantly increase the accuracy of the metric network in the latter stage. By comparing the existence of shortcut connections in InViNet, the low-level spatial feature is acknowledged as a useful complement for high-level semantic information. We use the following settings to train our network in two stages.

The feature extraction network is trained in the first stages. The branches in the feature extraction network are initialized with VGG16 trained weights by the ImageNet data set. Xavier[20] method initializes the new or modified layers. The low-level filters in VGG16 are acknowledged that they are beneficial for the shallow features, while higher-level features are more closely related to specific tasks. So, the learning rate multipliers in each layer are also set differently. The learning rate in each layer is the basic learning rate multiplied by its learning rate multiplier. The basic learning rate is 10^{-3}. The learning rate multiplier is 0.01 in Block1, Block2, and Block3. The learning rate multiplier is 0.05 in Block4 and Block5. The learning rate multiplier of FC6 and FC7 in VGG remains 1. Since all branches share weights, only one copy of the weights is in the feature extraction network. The optimizer uses the momentum SGD method. The momentum parameter is 0.9. Minibatch size in training is 16. The number of epochs is 2 000. The weight decay is 10^{-4}.

The metric network and shortcut connections are trained in the second stage. In metric network training, the weights trained well in the feature network are used as the initial value. The branches' weights slightly change during this training. Their learning rate multipliers are less than 10^{-2}. The basic learning rate is 10^{-3}. The weights are initialized with the Xavier method in new layers. Their learning rate multipliers are 1 in the metric network and shortcut layers. The number of epochs is adjusted to 2 500. The rest of the training parameters are the same as the first training.

All experiments run on a computer equipped with Nvidia TITAN XP GPU. Our experiment is implemented with Caffe.

2.3 Experimental result

To validate our approach, we have implemented the following experiments on different network architecture.

(1) Traditional method[9]. We enhance the object edges and use SURF to extract the infrared-visible image features. The similarity of images is measured by matching the feature points of the infrared and visible images.

(2) Baseline Network. MatchNet[13] is used as a baseline network. The Softmax loss function directly optimizes the whole network. There are no two phases in training. Two VGG16 branches train from scratch. The network has been over-fitting soon.

(3) MatchNet[13](F). MatchNet architecture improved with fine-tuning. Unlike the baseline network, its VGG16 branches are initialized by the weights trained with the ImageNet dataset.

(4) Pseudo-SiamNet[14] (F). The pseudo-Siamese Deep Compare Network architecture improved with fine-tuning. In the two VGG16 branches, the Conv1, Conv2 and Conv3 layers use their respective weights, whereas the Conv4 and Conv5 layers share the weights. The model weights also are initialized by the VGG16 trained to avoid over-fitting.

(5) InViNet (F+C). InViNet with fine-tuning and contrastive loss. We trained this network in two phases, which are described in Sec 2.2.

(6) InViNet (F+C+S). InViNet with fine-tuning, contrastive loss, and shortcut connection. The network adds shortcut connections.

(7) InViNet (F+T+S). InViNet with fine-tuning, triplet loss, and shortcut connection. This network is mainly to compare triplet loss and constrained loss.

The ROC curve usually measures a binary classification performance to avoid the imbalance between positive and negative samples. The commonly used evaluation metric is the false positive rate at 95% recall (Error@95%), the lower the better. Based on the experimental results, ROC curves are drawn for different methods. See Fig.6 for details.

From our experiments, the following conclusions can be summarized.

(1) In infrared-visible image patches matching, it is

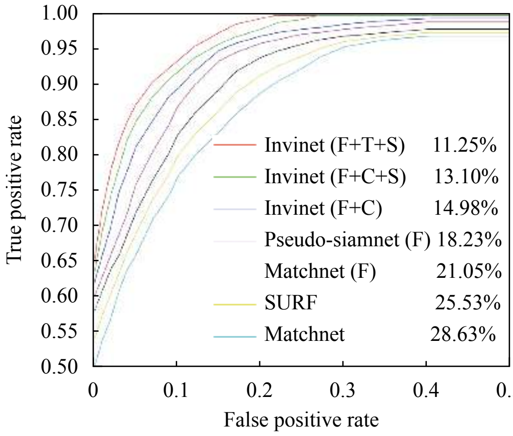


Fig.6 ROC curves for various methods. The numbers in the legends are FPR95 values. In the legend, the symbol “F” means the network uses fine-tuning with VGG16. The symbol “C” means that the contrastive loss is used in the extraction feature network. The symbol “T” means that the triplet loss is used in the extraction feature network. The symbol “S” means that shortcut connection is used

hard to extract common features in infrared and visible images with traditional methods due to the different imaging principles. The result is not satisfying.

(2) The few samples easily lead to over-fitting when the network is trained from scratch. With the fine-tuning, all deep learning networks show better performance than traditional algorithms. The fine-tuning can avoid over-fitting effectively.

(3) The pseudo-Siamese network performs better than the Siamese network. The explanation may be that the low-level convolution layers don ’t share weights in pseudo-Siamese networks. According to the different imaging principles of infrared and visible images, they can extract their unique shallow features from two separate branches.

To be concrete, we visualize the deep learned features of expression using t-SNE[21], a common tool used to visualize high-dimensional data. Our approach can effectively reduce the intra-class distance and enlarge the inter-class distance in Fig.7 which is beneficial for patches matching.

We show some top-ranking correct and incorrect results in InViNet in Fig.8. We find that incorrect results also may be easily mistaken by a human.

To further analyze our method results, we list the mean average precision (MAP) in the test set, which has five classes. The classes have never been used in the training process. As shown in Fig.9, our InViNet outperforms other approaches.

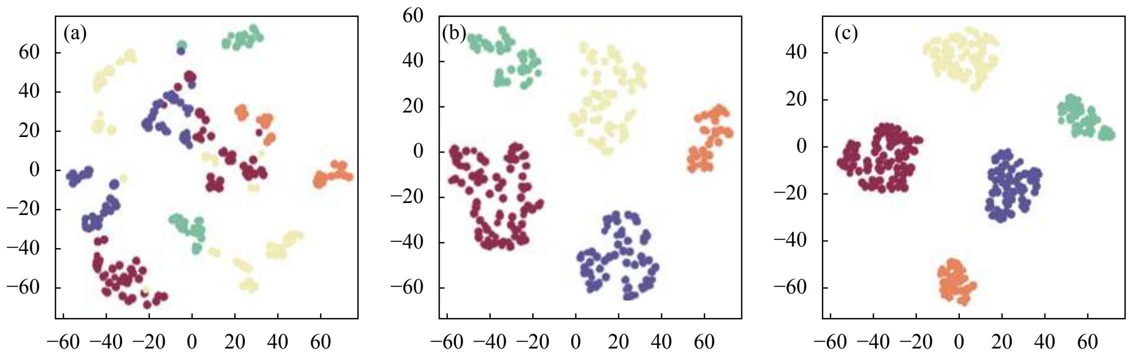


Fig.7 Visualization of the five class features in the test data set by the feature extraction network. (a) Features from the original network; (b) Features from the network with the contrastive loss; (c) Features from the network with the triplet loss

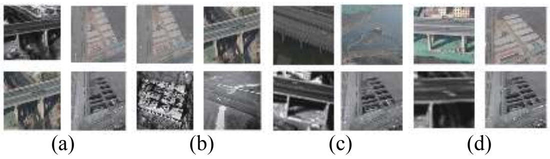


Fig.8 Top-ranking false and true results in overpass and factory image patches. (a) True positive samples; (b) True negative samples; (c) False positive samples; (d) False negative samples

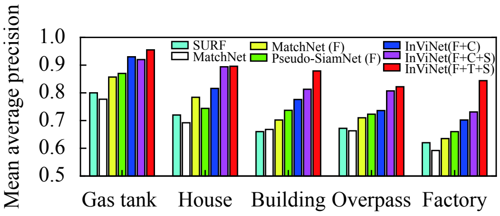


Fig.9 Performance matching in the test data set. In the legend, the symbols “F”, “C”, “T” and “S” have the same meaning in Fig. 6

3 Conclusions

Given the difficulty of infrared-visible image patches matching, this paper proposes an improved network based on deep learning. Compared to the previous method, our method can increase the accuracy from 78.95% to 88.75%. At present, it is difficult to obtain samples of visible and infrared images. There are many multi-sensor data sets available on the Internet. However, they are not fully utilized because there is no corresponding similar visible image. We believe that we can make full use of many multi-sensor images through unsupervised learning to further improve our matching performance in the future.

References:

[1] Yang Weiping, Shen Zhenkang. Matching technique and its application in aided inertial navigation [J]. ***Infrared and Laser Engineering***, 2007, 36(S2): 15-17. (in Chinese)

[2] Li Hongguang, Ding Wenrui, Cao Xianbin, et al. Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing [J]. ***Remote Sensing***, 2017, 9(5): 441.

[3] Wang Ning, Zhou Ming, Du Qinglei. A method for infrared visible image fusion and target recognition [J]. ***Journal of Air Force Early Warning Academy***, 2019, 33(5): 328-332.

[4] Mao Yuanhong, He Zhanzhuang, Ma Zhong. Infrared target classification with reconstruction transfer learning [J]. ***Journal of University of Electronic Science and Technology of China***, 2020, 49(4): 609-614. (in Chinese)

[5] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. ***International Journal of Computer Vision***, 2004, 60(2): 91-110.

[6] Bay H, Tuytelaars T, Gool L V. SURF: Speeded up robust features[C]//European Conference on Computer Vision, 2006, 3951: 404–417.

[7] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]//International Conference on Computer Vision, 2011: 2564-2571.

[8] Sima A A, Buckley S J. Optimizing SIFT for matching of short wave infrared and visible wavelength images [J]. ***Remote Sensing***, 2013, 5(5): 2037-2056.

[9] Li D M, Zhang J L. A improved infrared and visible images matching based on SURF [J]. ***Applied Mechanics and Materials***, 2013, 2418(651): 1637-1640.

[10] Chao Zhiguo, Wu Bo. Approach on scene matching based on histograms of oriented gradients [J]. ***Infrared and Laser Engineering***, 2012, 41(2): 513-516. (in Chinese)

[11] Cao Zhiguo, Yan Ruicheng, Song Jie. Approach on fuzzy shape context matching between infrared images and visible images [J]. ***Infrared and Laser Engineering***, 2008, 37(12): 1095-1100. (in Chinese)

[12] Jiao Anbo, Shao Liyun, Li Chenxi, et al. Automatic target recognition algorithm based on affine invariant feature of line grouping [J]. ***Infrared and Laser Engineering***, 2019, 48(S2): S226003. (in Chinese)

[13] Han X, Leung T, Jia Y, et al. MatchNet: Unifying feature and metric learning for patch-based matching[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3279-3286.

[14] Zagoruyko S, Komodakis N. Learning to compare image patches via convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 4353-4361.

[15] Hanif M S. Patch match networks: Improved two-channel and Siamese networks for image patch matching [J]. ***Pattern Recognition Letters***, 2019, 120: 54-61.

[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//ICLR 2015: International Conference on Learning Representations, 2015.

[17] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 1-9.

[18] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), 2006, 2: 1735-1742.

[19] Schroff F, Kalenichenko D, Philbin J. FaceNet: A unified embedding for face recognition and clustering[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 815-823.

[20] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]//Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010: 249-256.

[21] Van der Maaten L , Hinton G. Visualizing data using t-SNE [J]. ***Journal of Machine Learning Research***, 2008, 9(86): 2579-2605.