SSML-QNet: Scale-Separative Metric Learning Quadruplet Network for Multi-modal Image Patch Matching

Xiuwei Zhang[1], **Yi Sun**[2], **Yamin Han**[3*], **Yanping Li**[1], **Hanlin Yin**[1], **Yinghui Xing**[1], **Yanning Zhang**[1]

[1]School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an, China
[2]School of Cybersecurity, Northwestern Polytechnical University, Xi'an, China
[3]College of Information Engineering, Northwest A&F University, Yangling, China
wxzhang@nwpu.edu.cn, yaminhan@nwafu.edu.cn

Abstract

Multi-modal image matching is very challenging due to the significant diversities in visual appearance of different modal images. Typically, the existing well-performed methods mainly focus on learning invariant and discriminative features for measuring the relation between multi-modal image pairs. However, these methods often take the features as a whole and largely overlook the fact that different scale features for a same image pair may have different similarity, which may lead to sub-optimal results only. In this work, we propose a Scale-Separative Metric Learning Quadruplet network (SSML-QNet) for multi-modal image patch matching. Specifically, SSML-QNet can extract both relevant and irrelevant features of imaging modality with the proposed quadruplet network architecture. Then, the proposed Scale-Separative Metric Learning module separately encodes the similarity of different scale features with the pyramid structure. And for each scale, cross-modal consistent features are extracted and measured by coordinate and channel-wise attention sequentially. This makes our network robust to appearance divergence caused by different imaging mechanisms. Experiments on benchmark datasets (VIS-NIR, VIS-LWIR, Optical-SAR, and Brown) have verified that the proposed SSML-QNet is able to outperform other state-of-the-art methods. Furthermore, the cross-dataset transferring experiments on these four datasets also have shown that the proposed method has powerful ability of cross-dataset transferring.

1 Introduction

With the development of imaging sensor, more and more applications integrate multiple imaging sensors to perform the tasks with high performance requirements, such as military exploration, medical detection, and security monitoring [Barnea and Silverman, 1972]. As the key technology for these applications, multi-modal image matching has drawn increasing attention from the research community [Zagoruyko and Komodakis, 2015; Moreshet and Keller, 2021; Simo-Serra *et al.*, 2015; Savinov *et al.*, 2017], which aims to measure the similarities between two image patches. However, image matching, especially multi-modal image matching, is still an ill-posed problem that suffers from the significant diversities in visual appearance of different modal images.

Multi-modal methods in general have been extensively explored [Own and Hassanien, 2002]. At an early stage, almost traditional image matching algorithms were based on hand-designed feature descriptors, such as SIFT [Lowe, 2004], PCA-SIFT [Ke and Sukthankar, 2004], SURF [Bay *et al.*, 2006], and SSIF [Liu *et al.*, 2008], etc. The above methods have achieved good performance on single modal image matching task, but their accuracy and robustness were still limited due to the significant appearance divergence between different modal images.

Recently, the methods based on deep learning technology have achieved great progress in this field. They can generally fall into two groups: descriptor learning [Simo-Serra *et al.*, 2015; Balntas *et al.*, 2016a; Savinov *et al.*, 2017; Tian *et al.*, 2017; Mishchuk *et al.*, 2017; Quan *et al.*, 2019] and metric learning [Zagoruyko and Komodakis, 2015; Han *et al.*, 2015; Kumar BG *et al.*, 2016; Baruch and Keller, 2021; Moreshet and Keller, 2021]. Descriptor learning-based methods generate global descriptor of input image patches through deep neural network, and measure their similarity by simple descriptor distance, then distinguish matching and non-matching through a proper threshold. By contrast, metric learning-based methods adopt a metric network to convert image patch matching problem into a binary classification task (matching and non-matching), which consists of a feature extraction part and classification part. Compared to descriptor learning-based methods, metric learning-based methods are more flexible and effective, since they can simultaneously optimise feature representation and similarity measurement.

The existing metric learning-based methods have achieved a state-of-the-art performance in recent studies, but they only focus on learning invariant and discriminative features. Especially when measuring the similarity, they often take the features as a whole and largely overlook the fact that different scale features for a same image pair may have different sim-

*corresponding author

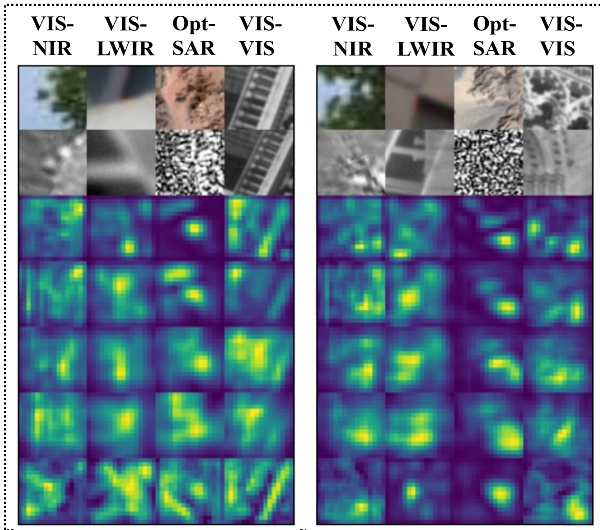

Figure 1: Visualization of similarity of different scale features. The left part shows the visualization result of matching patch-pairs, and the right part shows the visualization result of non-matching patch-pairs. From top to bottom are input image patch pairs, four feature map groups generated by the last four parallel convolutional layers of the SSML module, and the output feature map after fusion module, respectively. The brighter color area means that the corresponding feature response is stronger.

ilarity, which may lead to sub-optimal results only. To solve this problem, designing the fine-grained similarity metric is a promising solution. As motivated by the above analysis, we design a novel network architecture, which can encode the similarity of different scale features of the input image patch pair, as shown in Fig. 1.

The main contributions of this paper are summarized as follows:

1) A novel Scale-Separative Metric Learning Quadruplet network (SSML-QNet) is proposed for multi-model image patch matching. With a quadruplet network architecture, SSML-QNet can extract both relevant and irrelevant features of imaging modality. The proposed SSML (Scale-Separative Metric Learning) module separately encodes the similarity of different scale features. For each scale, SSML can accurately extract and measure cross-model consistent features by the operations of coordinate attention and Squeeze-and-Excitation (SE) attention. It makes our model robust to appearance divergence caused by different imaging mechanisms.

2) Experiments on benchmark datasets (VIS-NIR, VIS-LWIR, Optical-SAR, and Brown) have verified that the proposed SSML-QNet outperforms other state-of-the-art methods. The mean value of false positive rate at true positive rate equal to 95% (FPR95) is reduced to 0.75, 1.56, 0.65 and 0.58 on VIS-NIR, VIS-LWIR, Optical-SAR and Brown dataset, respectively. The transferring experiments also show that our method has powerful ability of cross-dataset transferring.

The remainder of this paper is organized as follows. The proposed method is described in Section 2. Section 3 presents the experiment configuration, experimental results and analysis. Finally, the conclusion is given in Section 4.

Layer	Output	Kernel	Stride	Pad	Dilation
Conv0 $\sim$ 1	$64\times64\times32$	3×3	1	1	1
Conv2 $\sim$ 3	$32\times32\times64$	3×3	1	1	1
Conv4	$32\times32\times128$	3×3	1	1	1
Conv5 $\sim$ 7	$16\times16\times128$	3×3	1	1	1

Table 1: The architecture of Siamese and Presudo-Siamese backbone.

2 Method

2.1 Overview

Fig. 2 shows the structure of the proposed Scale-Separative Metric Learning Quadruplet network, which is composed of three modules: quadruplet multi-model feature extraction module, scale-separative metric learning module, and multi-scale feature fusion and prediction module. When a new multi-modal image pair arrives, the quadruplet multi-model feature extraction module utilizes two types of CNN subnetworks to generate the relevant and irrelevant features of imaging modality. Then both relevant and irrelevant features are fed into the scale-separative metric learning module to encode the similarity of different scale features for increasing the accuracy of metric learning. After that, the multi-scale feature fusion and prediction module firstly fuses multi-scale features and the final prediction score is generated by adopting three fully connected layers. The technical details above are presented in the sections as below.

2.2 Quadruplet Multi-model Feature Extraction

Due to distinct imaging mechanisms, there are vast differences in visual appearance between different multi-model images. To better extract and represent similar features and discriminative features between image patch pairs, the quadruplet multi-model feature extraction module is adopted. As shown in the left part of Fig. 2, it contains four branches with the same structure. When a new multi-modal image pair arrives, a Siamese sub-network formed by the top two branches sharing parameters takes them as input to encode the features irrelevant to imaging modality. And the bottom two branches unsharing parameters form a Pseudo-Siamese sub-network to encode image pairs' features related to imaging modality. For each branch, it consists of six convolution layers, whose details are shown in Table 1. Specially, an instance normalization is added before batch normalization of the first three convolution layers, which reduces the feature difference caused by the illumination variation and different imaging mechanisms. Finally, the feature maps generated by both Siamese sub-network and Pseudo-Siamese sub-network are concatenated together and then used as inputs to the SSML module.

2.3 Scale-Separative Metric Learning Module

Multi-scale feature integration strategy and proper attention mechanism are proved to be beneficial for increasing the accuracy of metric learning [Hou *et al.*, 2021; Zhang *et al.*, 2021]. Inspired by this fact, we propose a Scale-Separative

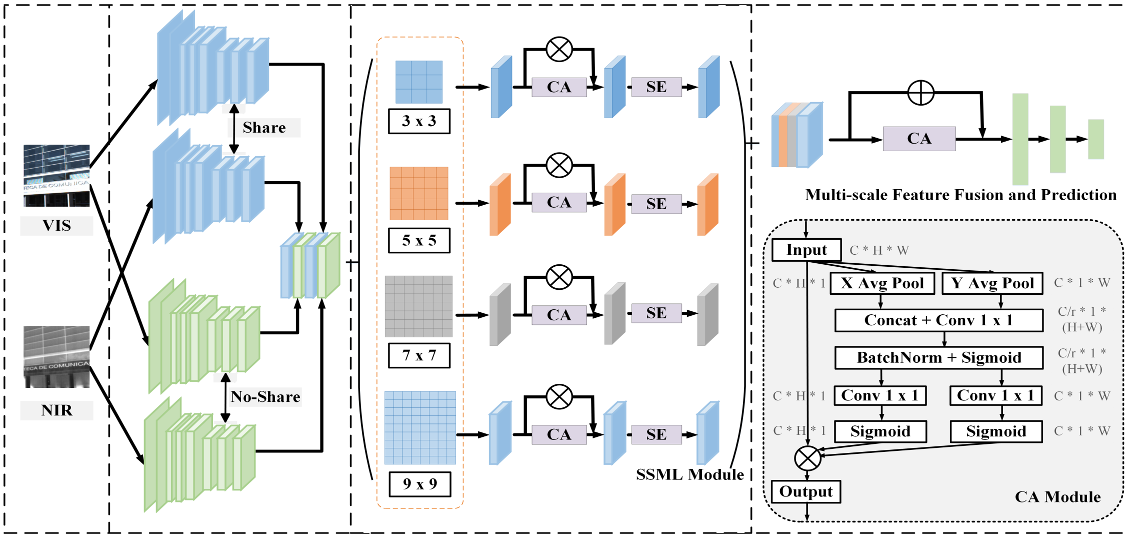

Figure 2: Overview of the proposed network architecture for multi-model image patch matching.

Metric Learning Module (SSML), which encodes the similarity of different scale features of the input image patch pair respectively, and then integrates them together to increase the accuracy of metric learning. As shown in Fig. 1, the SSML module focuses on the relevant features and suppresses irrelevant features of imaging modality. We do not add any additional supervisory information for this module. It is only learned by the objective function using a multi-scale feature encoder mechanism.

As illustrated in the middle part of Fig. 2, the SSML module mainly contains three steps. Firstly, four convolution layers with different receptive fields (3×3, 5×5, 7×7, and 9×9) are utilized to generate four feature map groups. By splitting the input feature maps into four groups with different scales, SSML module can better measure the similarity of each scale. Then, coordinate attention (CA) [Hou *et al.*, 2021] and Squeeze-and-Excitation attention (SE) is performed sequentially to encode coordination and channel-wise correlation for each feature map group. Finally, the four groups of feature map refined by CA and SE are regarded as the outputs of the SSML module. Visualization experimental result (Fig. 1) has shown that the similar features are highlighted for each scale by adopting the proposed SSML module, which makes the matching pairs and non-matching pairs becoming easier to be distinguished.

Specifically, given an input feature map $\mathbf{F} \in R^{L\times H\times W}$, the output of SSML module represented by $\mathbf{F}^{'} \in R^{L\times H\times W}$ can be computed as:

$$\begin{aligned} \mathbf{F}^{'} &= Concat(\mathbf{F}_{\mathbf{0}}^{'}, \mathbf{F}_{\mathbf{1}}^{'}, \mathbf{F}_{\mathbf{2}}^{'}, \mathbf{F}_{\mathbf{3}}^{'}) \\ \mathbf{F}_{\mathbf{i}}^{'} &= SE((CA(\mathbf{F}_{\mathbf{i}})) \otimes \mathbf{F}_{\mathbf{i}}) \\ \mathbf{F}_{\mathbf{0}}, \mathbf{F}_{\mathbf{1}}, \mathbf{F}_{\mathbf{2}}, \mathbf{F}_{\mathbf{3}} &= f^{3\times3}(\mathbf{F}), f^{5\times5}(\mathbf{F}), f^{7\times7}(\mathbf{F}), f^{9\times9}(\mathbf{F}) \end{aligned} \tag{1}$$

where $f^{n\times n}$ represents the convolution layer with kernel size of $n \times n$. $\mathbf{F}_{\mathbf{i}} \in R^{C\times H\times W}$ ($i = 1,2,3,4; C = L/4$) denotes one of four different scale feature maps. $\otimes$ is an element-wise multiplication. $CA(\cdot)$ represents the coordinate attention, $SE(\cdot)$ is the Squeeze-and-Excitation attention. $\mathbf{F}_{\mathbf{i}}^{'} \in R^{C\times H\times W}$ denotes scale-separative feature map.

The detailed structure of CA module presented in [Hou *et al.*, 2021] is illustrated in the right bottom of Fig. 2. Given an input $\mathbf{F}_{\mathbf{m}} \in R^{C\times H\times W}$, the spatial global average pooling kernel $(H, 1)$ and $(1, W)$ are performed along X-coordinate and Y-coordinate direction for each channel, respectively. Correspondingly, two feature maps with the size of $C \times H \times 1$ and $C \times 1 \times W$ are generated. Then, these two feature maps are concatenated by convolving with the filter (1×1) along the channel. After that, a new feature map with the size of $C/r \times 1 \times (H + W)$ is generate by performing Batch Normalization and Sigmoid operation. And through two sets of independent operations (a 1×1 convolution followed by a Sigmoid), the generated feature map is splited into two direction-aware attentions, i.e., Y-coordinate direction attention and X-coordinate direction attention with the size of $C \times H \times 1$ and $C \times 1 \times W$, respectively. Finally, these two attentions are multiplied with the input feature map $\mathbf{F}_{\mathbf{m}}$ to generate the final output feature map. Based on the above descriptions, it decomposes coordinate attention into two one-dimensional feature encoding processes. In this way, it can capture long-range dependencies along one spatial direction and preserve high precise location information along another spatial direction. Therefore, CA module can capture long-range dependencies with precise positional information.

Mathematically, $CA(\cdot)$ is defined as follows:

$$\begin{aligned} CA(\mathbf{F}_{\mathbf{m}}) &= F_m(c,i,j) \times T^h(c,i,1) \times T^w(c,1,j) \\ \mathbf{T}^{\mathbf{h}} = \delta(f^{1\times1}(\mathbf{F}_{\mathbf{t}}^{\mathbf{h}})), &\ \mathbf{T}^{\mathbf{w}} = \delta(f^{1\times1}(\mathbf{F}_{\mathbf{t}}^{\mathbf{w}})) \\ \mathbf{F}_{\mathbf{t}} = [\mathbf{F}_{\mathbf{t}}^{\mathbf{h}}, \mathbf{F}_{\mathbf{t}}^{\mathbf{w}}] &= split(\delta(\sigma(f^{1\times1}(Concat(\mathbf{G}^{\mathbf{h}}, \mathbf{G}^{\mathbf{w}}))))) \\ G^h(c,h) &= \frac{1}{W}\sum_{j=0}^{W-1} F_m(c,h,j) \end{aligned}$$

$$G^w(c,w) = \frac{1}{H}\sum_{i=0}^{H-1} F_m(c,i,w) \tag{2}$$

where $F_m(c,i,j)$ is the feature value of $\mathbf{F_m}$ at the position (c,i,j), in which c is the channel number, i and j denotes the X-coordinate and Y-coordinate, respectively. $f^{1\times 1}$ represents a 1×1 convolution layer. δ denotes Batch Normalization and σ is Sigmoid operator. $G^h(c,h)$ and $G^w(c,w)$ denote global pooling kernels to encode each channel along the X-coordinate and Y-coordinate direction, respectively. $\mathbf{G^h}$ and $\mathbf{G^w}$ denote the results after global pooling of the X-coordinate direction and Y-coordinate direction, respectively. $\mathbf{F_t} \in R^{C/r\times 1\times(H+W)}$ is the intermediate feature map that encodes spatial information in both X-coordinate and Y-coordinate direction. Then, $\mathbf{F_t}$ is split into two separate tensor along the spatial dimension, i.e. $\mathbf{F_t^h} \in R^{C/r\times H\times 1}$ and $\mathbf{F_t^w} \in R^{C/r\times 1\times W}$, r is a reduction ratio for controlling the feature map size. $\mathbf{T^h} \in R^{C\times H\times 1}$ and $\mathbf{T^w} \in R^{C\times 1\times W}$ denote two attention vectors of the Y-coordination and X-coordination direction, respectively.

Squeeze-and-Excitation can encode the relationship among feature channels by an attention vector, which is calculated among different channels of feature maps. The details about our implementation of SE model are as follows. The input feature maps firstly go through a global pooling layer, and output a vector with the same size as the number of input feature map channels. Then, a fully connected layer with 32 units followed by a ReLU activation function, a fully connected layer with C (the channel size of the input feature map) units, and a Sigmoid function are performed to generate the attention vector. Finally the input feature maps are weighted by the attention vector, and element-wise added with themselves to produce the channel-wise attentive features. Through the SE module, the feature maps contributing to the matching task are emphasized, and the others are restrained.

2.4 Multi-scale Feature Fusion and Prediction

Multi-scale feature fusion is very important for accurate prediction, which is denoted $MFFP$. Given the output feature map of SSML module $\mathbf{F}^{'}$, we first carry out a coordinate attention operation on $\mathbf{F}^{'}$. Then, an addition operation is performed to add $\mathbf{F}^{'}$ and CA attentive features together. After that, a 3×3 convolution is performed. Finally, three fully connected layers are adopted to predict the result. $MFFP(\cdot)$ can be described by the following formula.

$$MFFP(\mathbf{F}^{'}) = FC_2(FC_{128}(FC_{512}(f^{3\times 3}(CA(\mathbf{F}^{'})\oplus\mathbf{F}^{'})))) \tag{3}$$

where $CA(\cdot)$ is the coordinate attention as shown in Formula 2. $\mathbf{F}^{'}$ is the output feature map by SSML module. $\oplus$ is an element-wise addition operator. FC_2, FC_{128}, FC_{512} represent fully connected layers with 2, 128, and 512 units, respectively.

2.5 Loss Function

The image patch matching can be considered as a binary classification task (matching and non-matching). Cross-entropy loss is commonly used in classification task. In this paper, we adopt cross-entropy loss to train the network. In fact, we also considered other loss functions, including contrastive loss, hingle loss, and focal loss for experiments. However, these loss functions did not outperform cross entropy loss. Therefore, cross entropy loss $L_{en}$is more suitable for binary tasks like our image patch matching.

$$L_{en} = ylog\hat{y} + (1-y)log(1-\hat{y}) \tag{4}$$

where y and $\hat{y}$ represent the ground truth and the predictive value, respectively.

3 Experiments

3.1 Datasets

To verify the effectiveness of the proposed method, we carry out experiments on public multi-model image datasets, including VIS-NIR, VIS-LWIR and Optical-SAR, as well as a single spectral multi-view stereo correspondence dataset named Brown.

1) VIS-NIR is a multi-modal image patch matching dataset [Brown and Süsstrunk, 2011], which consists of more than 1,000,000 image patch pairs of visual spectrum and near-infrared spectrum. One Half of these patch pairs are matching pairs and the other half are non-matching pairs. The size of each image patch is 64×64 pixels. Totally, there are nine categories in this dataset, i.e., Country, Field, Forest, Indoor, Mountain, Oldbuilding, Street, Urban and Water. Same as the methods [Baruch and Keller, 2021; Quan *et al.*, 2021], the proposed model is trained on the Country category and test on the other categories. Since great differences among these categories, it is hard to obtain a good generalization performance on all test categories.

2) VIS-LWIR is a multi-modal dataset of visual spectrum (VIS) and long-wave infrared (LWIR) spectrum [Aguilera *et al.*, 2015]. It contains 44 VIS-LWIR image pairs, which are strictly aligned in time and space. Similar to VIS-NIR dataset, we also crop image patch pairs from VIS and LWIR images centered on SIFT points. The patch size is 64×64 pixels. Following previous studies [Quan *et al.*, 2021], our method is also trained on one half of the patch pairs and tested on the other half. There are significant appearance differences between VIS image and its corresponding LWIR image.

3) Optical-SAR is a multi-modal image patch dataset, which contains optical images and synthetic aperture radar (SAR) images. We generate these image patch pairs in the same way as VIS-NIR dataset. SEN1-2 [Schmitt *et al.*, 2019] dataset contains 282,384 image pairs of optical images and corresponding SAR. Similar to [Quan *et al.*, 2021], 583,180 image patch pairs are utilized for training and the other 248274 pairs for testing.

4) Brown is a single spectral multi-view stereo correspondence dataset [Brown *et al.*, 2010]. It contains three subsets: Liberty, Notredame and Yosemite, which contains 100K, 200K, and 500K image patch pairs, respectively. For each subset, one half of the patch pairs are matching pairs with the same 3D point and the other half are non-matching pairs. The patch size is 64×64 pixels. Like the methods [Tian *et al.*, 2017; Han *et al.*, 2015; Zagoruyko and Komodakis, 2015],

Method	Field	Forest	Indoor	Mountain	Oldbuilding	Street	Urban	Water	Mean
SIFT [Lowe, 2004]	39.44	11.39	10.13	28.63	19.69	31.14	10.85	40.33	23.95
GISIFT [Firmenichy *et al.*, 2011]	34.75	16.63	10.63	19.52	12.54	21.80	7.21	25.78	18.60
EHD [Aguilera *et al.*, 2012]	33.85	19.61	24.23	26.32	17.11	22.31	3.77	19.80	20.87
LGHD [Shechtman and Irani, 2007]	16.52	3.78	7.91	10.66	7.91	6.55	7.21	12.76	9.16
PN-Net [Balntas *et al.*, 2016a]	20.09	3.27	6.36	11.53	5.19	5.62	3.31	10.72	8.26
Q-Net [Savinov *et al.*, 2017]	17.01	2.70	6.16	9.61	4.61	3.99	2.83	8.44	6.91
L2-Net [Tian *et al.*, 2017]	16.77	0.76	2.07	5.98	1.89	2.83	0.62	11.11	5.25
HardNet [Mishchuk *et al.*, 2017]	10.89	0.22	1.87	3.09	1.32	1.30	1.19	2.54	2.80
Siamese [Simo-Serra *et al.*, 2015]	15.79	10.76	11.60	11.15	5.27	7.51	4.60	10.21	9.61
Pseudo-Siamese [Zagoruyko and Komodakis, 2015]	17.01	9.82	11.17	11.86	6.75	8.25	5.65	12.04	10.31
2-Channel [Zagoruyko and Komodakis, 2015]	9.96	0.12	4.40	8.89	2.30	2.18	1.58	6.40	4.47
SCFDM [Quan *et al.*, 2018]	7.91	0.87	3.93	5.07	2.27	2.22	0.85	4.75	3.48
Hybrid [Baruch and Keller, 2021]	5.62	0.53	3.58	3.51	2.23	1.82	1.90	3.05	2.52
Moreshet & K+ [Moreshet and Keller, 2021]	4.22	0.13	1.48	1.03	1.06	1.03	0.9	1.9	1.44
Quan & W+ [Quan *et al.*, 2021]	4.21	0.11	1.12	0.87	0.67	0.56	0.43	1.90	1.23
AFD-Net [Quan *et al.*, 2019]	3.47	0.08	1.48	0.68	0.71	0.42	0.29	1.48	1.08
MFD-Net [Yu *et al.*, 2022]	2.59	**0.02**	1.24	0.95	**0.48**	**0.24**	**0.12**	**1.44**	0.88
SSML-QNet	**0.97**	0.55	**0.65**	**0.24**	0.62	0.69	0.43	1.71	**0.73**

Table 2: Comparisons with the-state-of-the-art on the VIS-NIR dataset.

Test Dataset / Train Dataset	Brown			Mean
	Notredame	Yosemite	Liberty	
VIS-NIR	1.55	2.63	2.26	2.15
VIS-LWIR	2.92	3.16	2.50	2.86
Optical-SAR	3.51	2.83	4.69	3.68

Table 3: Cross-dataset Transfering Performance: trained on other datasets and test on Brown dataset.

Dataset / Method	VIS-LWIR	Optical-SAR
Siamese	42.62	17.56
Pseudo-Siamese	43.27	19.30
2Channel	22.95	7.35
Hybrid	18.09	14.90
SSML-QNet	1.56	0.65

Table 4: Comparisons with the-state-of-the-art on the VIS-LWIR dataset and Optical-SAR dataset.

Test Dataset / Train Dataset		VIS-NIR	VIS-LWIR	Optical-SAR
Brown	Yosemite	1.91	6.05	6.57
	Notredame	1.21	3.95	2.49
	Liberty	1.57	3.96	1.67
VIS-NIR		-	6.72	10.44
VIS-LWIR		1.86	-	2.36
Optical-SAR		7.96	18.01	-

Table 5: Cross-dataset Transferring Performance: trained on other datasets and test on the VIS-NIR, VIS-LWIR and Optical-SAR dataset.

the proposed model is trained on one of three subsets and test on the other subsets.

3.2 Implementation Details

The code of the proposed model is implemented by Pytorch. It is trained with Adam optimizer, and the learning rate is 0.0001. The batch size is set to 128. The training time is set to 80 epochs, the momentum is initially set to 0.9 with the decay factor 0.9. The cross-entropy loss is adopted to train the network. To quantitatively evaluate the matching performance, the false positive rate at true positive rate (positive recall) equal to 95% (FPR95) is adopted.

3.3 Comparison with the State-of-the-Arts

1) Results on VIS-NIR Dataset: The proposed method is compared with the state-of-the-art image patch matching methods on the VIS-NIR dataset. Totally, there are seventeen comparison algorithms. Among these methods, SIFT [Lowe, 2004], GISIFT [Firmenichy *et al.*, 2011], EHD [Aguilera *et al.*, 2012], LGHD [Shechtman and Irani, 2007] are traditional hand-designed descriptor-based methods, which are limited by human prior-knowledge and have poor robustness and adaptability. PN-NET [Balntas *et al.*, 2016a], L2-Net [Tian *et al.*, 2017], and HardNet [Mishchuk *et al.*, 2017] are descriptor learning-based methods. They focus on learning a representation that can enable the two matched features as close as possible, while making non-matched features far apart. Siamese [Simo-Serra *et al.*, 2015], Pseudo-Siamese [Zagoruyko and Komodakis, 2015], 2-Channel [Zagoruyko and Komodakis, 2015], SCFDM [Quan *et al.*, 2018], Hybrid [Baruch and Keller, 2021], Moreshet & K+ [Moreshet and Keller, 2021], Quan & W+ [Quan *et al.*, 2021], AFD-Net [Quan *et al.*, 2019], and MFD-Net [Yu *et al.*, 2022] are all metric learning-based methods. As shown in Table 2, our method outperforms other comparison methods. Compared with the second-best method MFD-Net [Yu *et al.*, 2022], the mean FPR95 value of our method is reduced by 0.15. Compared with Hybrid [Baruch and Keller, 2021], which is similar to our baseline and also has both Siamese sub-network and Pseudo-Siamese sub-network, the mean FPR95 value of our method is reduced by 1.79. It demonstrates that our method can effectively extract and measure the similarity of multi-model image patches by using the proposed SSML module and fusion strategy.

Method \ Test Dataset	Field	Forest	Indoor	Mountain	Oldbuilding	Street	Urban	Water	Mean
Concat	1.13	0.70	0.68	0.26	0.69	0.81	0.48	1.82	0.82
Sum	2.23	2.51	1.80	1.77	2.41	1.69	1.94	2.53	2.11
MSSAM	1.88	1.17	1.01	0.50	0.98	1.22	0.68	2.77	1.28
TF	1.89	1.12	0.98	0.41	0.88	1.10	0.64	2.71	1.22
CA(Ours)	**0.97**	**0.55**	**0.65**	**0.24**	**0.62**	**0.69**	**0.43**	**1.71**	**0.73**

Table 6: Fusion strategy comparison results on the VIS-NIR dataset.

Training	Notredame	Yosemite	Liberty	Yosemite	Liberty	Notredame	Mean
Test	Liberty		Notredame		Yosemite		
RootSIFT [Arandjelovic, 2012]	29.65		22.06		26.71		26.14
L-BGM [Trzcinski *et al.*, 2012]	18.05	21.03	14.15	13.73	19.63	15.86	17.08
Convex optimization [Simonyan *et al.*, 2014]	12.42	14.58	7.22	6.82	11.18	10.08	10.38
TNet-TGLoss [Kumar BG *et al.*, 2016]	9.91	13.45	3.91	5.43	10.65	9.47	8.80
SNet-GLoss [Kumar BG *et al.*, 2016]	6.39	8.43	1.84	2.83	6.61	5.57	5.27
PN-Net [Balntas *et al.*, 2016a]	8.13	9.65	3.71	4.23	8.99	7.21	6.98
Q-Net [Savinov *et al.*, 2017]	7.64	10.22	4.07	3.76	9.34	7.69	7.12
DeepDesc [Simo-Serra *et al.*, 2015]	10.90		4.40		5.69		6.99
TFeat-ration [Balntas *et al.*, 2016b]	8.07	9.53	3.47	4.23	8.53	7.24	6.84
TFeat-margin [Balntas *et al.*, 2016b]	7.22	9.79	3.12	3.85	7.82	7.08	6.47
L2-Net [Tian *et al.*, 2017]	2.36	4.70	0.72	1.29	2.57	1.71	2.22
HardNet [Mishchuk *et al.*, 2017]	1.49	2.51	0.53	0.78	1.96	1.84	1.51
MathchNet [Han *et al.*, 2015]	6.90	10.77	3.87	5.67	10.88	8.39	7.44
DeepCompare [Zagoruyko and Komodakis, 2015]	4.85	7.20	1.90	2.11	5.00	4.10	4.19
SCFDM [Quan *et al.*, 2018]	1.47	4.54	1.29	1.96	2.91	5.20	2.89
Quan & W+ [Quan *et al.*, 2021]	1.47	2.09	0.50	0.77	1.69	1.75	1.38
Moreshet & K+ [Moreshet and Keller, 2021]	**0.35**	0.91	1.31	0.85	1.58	0.41	0.9
AFD-Net [Quan *et al.*, 2019]	1.53	2.31	0.47	0.72	1.63	1.88	1.42
MFD-Net [Yu *et al.*, 2022]	1.21	2.10	**0.40**	0.74	1.85	1.77	1.35
SSML-QNet	0.85	**0.86**	0.53	**0.65**	**0.47**	**0.12**	**0.58**

Table 7: Comparisons with the-state-of-the-art on the Brown dataset.

2) Results on VIS-LWIR Dataset: As shown in Table 4, we compare the proposed method with five state-of-the-art methods on VIS-LWIR dataset, including Siamese [Simo-Serra *et al.*, 2015], Pseudo-Siamese [Zagoruyko and Komodakis, 2015], 2-Channel [Zagoruyko and Komodakis, 2015], and Hybrid [Baruch and Keller, 2021]. Our proposed method achieves an excellent performance. The mean FPR95 value of our method is 1.56.

3) Results on Optical-SAR Dataset: As shown in Table 4, although there are significant appearance differences between optical images and SAR images, the mean FPR95 value of our method is 0.65. Compared with the other methods, the performance of our method is very excellent.

4) Results on Brown Dataset: To demonstrate the generalization ability of the proposal, we also test and compare SSML-QNet with other methods on Brown, i.e., a single spectral multi-view stereo correspondence benchmark dataset. As shown in Table 7, compared with the second-best method Moreshet & K+ [Moreshet and Keller, 2021], the mean FPR95 value is significantly improved by 0.32. This improvement demonstrates that our method can effectively encode and evaluate the similarity between images of different views and has a better generalization ability.

From the above four experiments, the proposed method achieves much better performance than the other methods. It can demonstrate that our model is effective not only for multi-model images, but also for single-modal images. Note that compared with VIS-NIR and Optical-SAR datasets, the mean FPR95 value on VIS-LWIR dataset is lower. One possible reason is that the correspondence between VIS image and LWIR image is more diverse, since the thermal radiation energy of observed objects will vary with many factors, such as object status, material quality, environment temperature, and observation distance.

3.4 Ablation Study

To verify the effectiveness of each module, we conduct ablation experiments. "BL" means our baseline, consisting of only quadruplet multi-model feature extraction and fully connected layers. "SSML" is our scale-separative metric learning module. "CA" means adopting coordinate attention in the feature fusion stage. "Sia" and "Pse-Sia" represents only considering Siamese sub-network and Pseudo-Siamese sub-network of our baseline, respectively.

As shown in Table 8, by adding SSML and CA into our baseline respectively, the mean FPR95 value is reduced by 1.28 and 0.23. By adding both of SSML and CA modules, the improvement becomes more significant, reaching 1.35. It can

BL	SSML	CA	Sia	Pse-Sia	Field	Forest	Indoor	Mountain	Oldbuilding	Street	Urban	Water	Mean
✓					2.91	2.30	1.84	0.82	1.73	2.04	1.34	3.84	2.10
✓	✓				1.13	0.70	0.68	0.26	0.69	0.81	0.48	1.82	0.82
✓		✓			2.59	2.02	1.50	0.75	1.50	1.84	1.28	3.45	1.87
	✓	✓	✓		1.99	1.22	0.78	0.46	0.88	1.27	0.66	2.69	1.24
	✓	✓		✓	2.62	1.87	1.69	0.80	1.23	1.56	1.08	3.60	1.81
✓	✓	✓			**0.97**	**0.55**	**0.65**	**0.24**	**0.62**	**0.69**	**0.43**	**1.71**	**0.73**

Table 8: Ablation results evaluated on the VIS-NIR dataset.

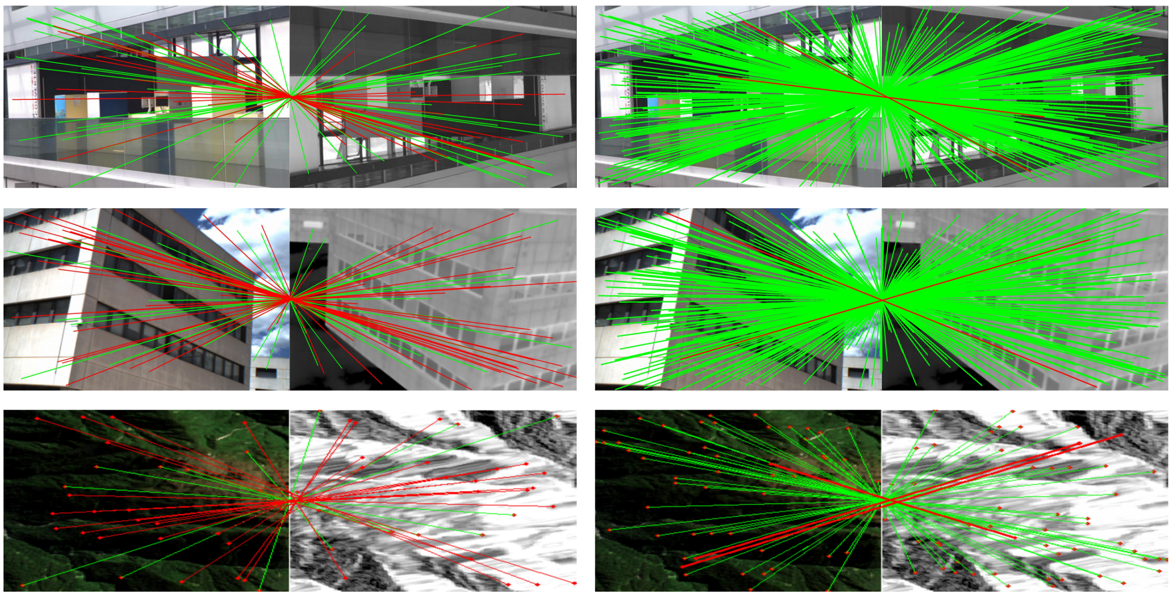

Figure 3: Image matching visualization result, from top to bottom: VIS-NIR, VIS-LWIR, Optical-SAR.

demonstrate the effectiveness of the proposed SSML module and the fusion way by using CA. When only considering Siamese or Pseudo-Siamese sub-network, the mean FPR95 value is 1.24 and 1.81, respectively. While considering both of them, the mean FPR95 value is significantly reduced to 0.73. Therefore, our quadruplet network structure is effective for multi-model image matching task.

3.5 Fusion Strategy Comparison

To verify the effect of our fusion strategy (CA), four typical fusion ways are adopted to compare with our method. These four comparison fusion ways are simple concatenation, element-wise sum fusion, multi-scale spatial feature attention module (MSSAM) [Zhang *et al.*, 2022], and Transformer encode module, which are denoted as “Concat”, “Sum”, “MSSAM” and “TF”, respectively. Similar to CA, Transformer encode module can also establish long-distance dependencies and obtain global context information. MSSAM can automatically learn the weight map of each scale feature group and effectively fuse the spatial detail information of different scale feature groups. We conduct an experiment to replace CA by other comparison fusion way based on our SSML-QNet, respectively. Specifically, the Transformer encoder has two layers and each layer consists of two multi-head attention blocks. Same as Moreshet et al. [Moreshet and Keller, 2021], the VIT pretrained model is loaded for Transformer encoder To improve the image matching performance. The experiment results in Table 6 show that compared with Transformer encoder and MSSAM, the mean FPR95 value of our method is improved by 0.49 and 0.55 by using CA, respectively. Therefore, CA is more suitable for fusing the features extracted by SSML module.

3.6 Cross-dataset Transferring Performance

To evaluate the cross-dataset transferring performance of the proposed method, we select one dataset for testing, and adopt other three datasets to train three models, respectively. The experimental results are shown in Table 3 and Table 5. We can see that except the model trained on Optical-SAR dataset, the other models can achieve good cross-dataset transferring performance. The possible reason is that the imaging mechanism and features of SAR images are far apart from other modal images and our model effectively learn the features related to imaging modality through quadruplet multi-model feature extraction module. While, the models trained on Liberty of Brown and VIS-LWIR performs better on Optical-SAR dataset and gains the mean FPR95 value of 1.67 and 2.36, respectively. Experimental results demonstrate that our

Method	FPS	Memory(MB)	FPR95
Siamese	1453.85	2647	9.61
Pseudo-Siamese	1359.34	2731	10.31
2Channel	1379.00	2283	4.47
Hybrid	1442.30	2947	2.52
SSML-QNet	1384.97	3013	0.73

Table 9: The comparison results in terms of computation efficiency, memory usage and matching performance.

network has good generalization performance and robustness.

3.7 Computational Efficiency and Memory Usage

Table 9 shows the comparison of computation efficiency, memory usage and matching performance. All compared methods are tested on the same workstation (one RTX 3090Ti). Our method achieves the best FPR95 score while achieving competitive computational efficiency.

3.8 Image Matching Visualization Experiment

This section analyzes the visualization results based on the matching point pairs learned from the proposed SSML-QNet. We compare the proposed method with baseline method on three multi-modal dataset. As shown in Fig. 3, our method achieves an excellent performance. The visualization results of the baseline model are illustrated in Fig. 3(a), and that of the proposed SSML-QNet model are shown in Fig. 3(b). All modal images are processed by geometric transformation (rotation = 180°, translation = 2 pixels). The green lines represent matches and the red lines denote non-matches. The experimental results show that the proposed SSML-QNet achieves good results in VIS-NIR and VIS-LWIR image pairs, but achieves relatively less matching point pairs in SAR. There are two possible reasons. Firstly, the imaging mechanism and characteristics of SAR images are quite different from those of other modal images, which leads to poor generalization effect. Secondly, it may be difficult to detect more robust feature points in SAR images due to the influence of traditional detection operators in the early feature point detection.

4 Conclusion

In this paper, we proposed a scale-separative metric learning quadruplet network for multi-modal image patch matching, named SSML-QNet. It can effectively extract cross domain consistent features and measure feature similarity. The experiments show that our proposal method performs much better than the-state-of-the-art methods on three multi-modal datasets (VIS-NIR, VIS-LWIR and Optical-SAR) and a single modal Brown dataset, and also has excellent cross-dataset transferring performance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants U19B2037, 61971356, 62201473 and 62201467, by the National Natural Science Foundation of Shaanxi province under Grant 2021KWZ-03 and 2022JQ-686. Thank the providers of the dataset.

References

[Aguilera *et al.*, 2012] Cristhian Aguilera, Fernando Barrera, Felipe Lumbreras, Angel D Sappa, and Ricardo Toledo. Multispectral image feature points. *Sensors*, 12(9):12661–12672, 2012.

[Aguilera *et al.*, 2015] Cristhian A Aguilera, Angel D Sappa, and Ricardo Toledo. Lghd: A feature descriptor for matching across non-linear intensity variations. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 178–181. IEEE, 2015.

[Arandjelovic, 2012] R. Arandjelovic. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[Balntas *et al.*, 2016a] Vassileios Balntas, Edward Johns, Lilian Tang, and Krystian Mikolajczyk. Pn-net: Conjoined triple deep network for learning local image descriptors. *arXiv preprint arXiv:1601.05030*, 2016.

[Balntas *et al.*, 2016b] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

[Barnea and Silverman, 1972] Daniel I Barnea and Harvey F Silverman. A class of algorithms for fast digital image registration. *IEEE transactions on Computers*, 100(2):179–186, 1972.

[Baruch and Keller, 2021] Elad Ben Baruch and Yosi Keller. Joint detection and matching of feature points in multimodal images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Bay *et al.*, 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[Brown and Süsstrunk, 2011] Matthew Brown and Sabine Süsstrunk. Multi-spectral sift for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE, 2011.

[Brown *et al.*, 2010] Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):43–57, 2010.

[Firmenichy *et al.*, 2011] Damien Firmenichy, Matthew Brown, and Sabine Süsstrunk. Multispectral interest points for rgb-nir image registration. In *2011 18th IEEE international conference on image processing*, pages 181–184. IEEE, 2011.

[Han *et al.*, 2015] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3279–3286, 2015.

[Hou *et al.*, 2021] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13713–13722, 2021.

[Ke and Sukthankar, 2004] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[Kumar BG *et al.*, 2016] Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5385–5394, 2016.

[Liu *et al.*, 2008] Li Liu, FY Peng, Kun Zhao, and YP Wan. Simplified sift algorithm for fast image matching. *Infrared and Laser Engineering*, 37(1):181–184, 2008.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Mishchuk *et al.*, 2017] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in neural information processing systems*, 30, 2017.

[Moreshet and Keller, 2021] Aviad Moreshet and Yosi Keller. Paying attention to multiscale feature maps in multimodal image matching. *arXiv preprint arXiv:2103.11247*, 2021.

[Own and Hassanien, 2002] Hala S Own and Aboul Ella Hassanien. Multiresolution image registration algorithm in wavelet transform domain. In *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No. 02TH8628)*, volume 2, pages 889–892. IEEE, 2002.

[Quan *et al.*, 2018] Dou Quan, Shuai Fang, Xuefeng Liang, Shuang Wang, and Licheng Jiao. Cross-spectral image patch matching by learning features of the spatially connected patches in a shared space. In *Asian Conference on Computer Vision*, pages 115–130. Springer, 2018.

[Quan *et al.*, 2019] Dou Quan, Xuefeng Liang, Shuang Wang, Shaowei Wei, Yanfeng Li, Ning Huyan, and Licheng Jiao. Afd-net: Aggregated feature difference learning for cross-spectral image patch matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3017–3026, 2019.

[Quan *et al.*, 2021] Dou Quan, Shuang Wang, Yi Li, Bowu Yang, Ning Huyan, Jocelyn Chanussot, Biao Hou, and Licheng Jiao. Multi-relation attention network for image patch matching. *IEEE Transactions on Image Processing*, 30:7127–7142, 2021.

[Savinov *et al.*, 2017] Nikolay Savinov, Akihito Seki, Lubor Ladicky, Torsten Sattler, and Marc Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1822–1830, 2017.

[Schmitt *et al.*, 2019] Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. Sen12ms–a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. *arXiv preprint arXiv:1906.07789*, 2019.

[Shechtman and Irani, 2007] Eli Shechtman and Michal Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[Simo-Serra *et al.*, 2015] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international conference on computer vision*, pages 118–126, 2015.

[Simonyan *et al.*, 2014] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1573–1585, 2014.

[Tian *et al.*, 2017] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 661–669, 2017.

[Trzcinski *et al.*, 2012] T. Trzcinski, Mario Christoudias, V. Lepetit, and Pascal V Fua. Learning image descriptors with the boosting-trick. In *Neural Information Processing Systems*, 2012.

[Yu *et al.*, 2022] Chuang Yu, Yunpeng Liu, Chenxi Li, Lin Qi, Xin Xia, Tianci Liu, and Zhuhua Hu. Multi-branch feature difference learning network for cross-spectral image patch matching. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[Zagoruyko and Komodakis, 2015] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015.

[Zhang *et al.*, 2021] Hu Zhang, Keke Zu, Jian Lu, Yuru Zou, and Deyu Meng. Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. *arXiv preprint arXiv:2105.14447*, 2021.

[Zhang *et al.*, 2022] Xiuwei Zhang, Mu Tian, Yinghui Xing, Yuanzeng Yue, Yanping Li, Hanlin Yin, Runliang Xia, Jin Jin, and Yanning Zhang. Adhr-cdnet: Attentived differential high-resolution change detection network for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.