# AFD-Net: Aggregated Feature Difference Learning for Cross-Spectral Image Patch Matching

Dou Quan[1] Xuefeng Liang[1,2] Shuang Wang[1] Shaowei Wei[1] Yanfeng Li[1]
Ning Huyan[1] Licheng Jiao[1]
[1]School of Artificial Intelligence, Xidian University, Shaanxi, China [2]Kyoto University, Kyoto, Japan
shwang@mail.xidian.edu.cn

## Abstract

*Image patch matching across different spectral domains is more challenging than in a single spectral domain. We consider the reason is twofold: 1. the weaker discriminative feature learned by conventional methods; 2. the significant appearance difference between two images domains. To tackle these problems, we propose an aggregated feature difference learning network (AFD-Net). Unlike other methods that merely rely on the high-level features, we find the feature differences in other levels also provide useful learning information. Thus, the multi-level feature differences are aggregated to enhance the discrimination. To make features invariant across different domains, we introduce a domain invariant feature extraction network based on instance normalization (IN). In order to optimize the AFD-Net, we borrow the large margin cosine loss which can minimize intra-class distance and maximize inter-class distance between matching and non-matching samples. Extensive experiments show that AFD-Net largely outperforms the state-of-the-arts on the cross-spectral dataset, meanwhile, demonstrates a considerable generalizability on a single spectral dataset.*

## 1. Introduction

Establishing the local correspondences between images plays a crucial role in many computer vision tasks, *e.g.* image retrieval [19], multi-view stereo reconstruction [29], and image registration [35]. Recently, increasing attention has been focused on the cross-spectral image matching because the different spectral domains provide complementary information [10, 15, 24]. For example, visible spectrum images (VIS) and near-infrared images (NIR) can mutually compensate the rich color information and the high texture structure [2]. Therefore, matching images across different domains becomes a new challenge.

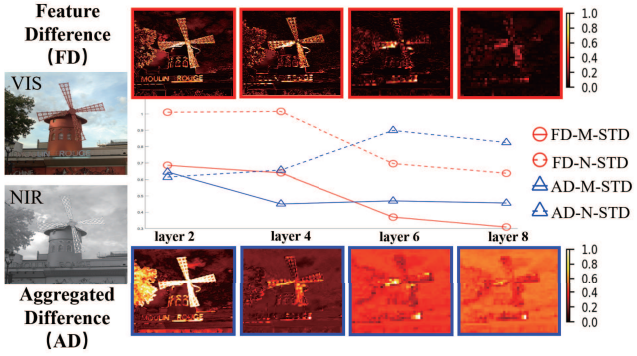



Figure 1. The changes of feature difference (FD) and aggregated difference (AD), and their standard deviations (STD) at different layers for matching (M) and non-matching (N) samples in a cross-spectral dataset.

The conventional matching methods are based on the handcraft local feature descriptors, such as SIFT [21], SURF [6], GISIFT [10] and shape context [7]. They perform rather well on the visible light images. However, as shown in Fig. 1, cross-spectral images appear significantly different at pixel-level due to the varied imaging mechanisms, which severely degrades the performance of handcraft features in the matching task. Recently, the deep learning-based methods have shown unprecedented advantage in feature learning for image matching. Generally, there are two major categories: *Descriptor learning* [3, 5, 16, 22, 30, 32] and *Metric learning* [2, 11, 25, 38]. Descriptor learning methods extract the high-level features of input image patches through the convolutional network, and measure their similarity by feature distance. Instead, metric learning methods transform this problem into a binary classification task (matching and non-matching) by adding a classifier network after the feature extraction network. Commonly, the framework is optimized by the cross-entropy loss. One can see that both of these methods merely rely on the high-level features, because they are more abstract and invariant

This work is supported by the National Natural Science Foundation of China (No.61771379), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No.61621005), the Fundamental Research Funds of the Central Universities of China (No.JC1904) and the Program for Cheung Kong Scholars and Innovative Research Team in University (No.IRT_15R53).

to rotation, perspective and scales [17, 27]. We find features in other levels are also useful due to involving more texture information, and the feature difference (FD) of patch-pairs can contribute to the matching prediction. The reason is that FD could cancel the same signal of matching samples, but amplify the different signal of non-matching samples. Unfortunately, this rich information has not been used for image patch matching yet.

In Fig. 1, we use one sample pair and a set of 60K samples to illustrate the feature difference (FD). The upper maps represent FD of the sample. The curves in the middle denote the standard deviations (STD) of the set of image patch-pairs at each layer. In general, a large STD denotes rich information. Figure 1 shows the FDs are decreasing for both matching (red line) and non-matching (red dot line) samples when going deeper into the network. Thus, the high-level features may not be sufficient to discriminate the hard samples. Instead, we aggregate the neighbor FDs and process them sequentially through convolution operators, and name them as Aggregated Difference (AD). The lower maps represent AD of the sample. Interestingly, the STD of ADs of matching samples (blue line) decreases initially and tends to stable in deeper layers. On the contrary, that of non-matching samples (blue dot line) appears a notable increase. This observation inspires us that the AD aggregates more lower level information and brings more discriminative learning signal for feature learning. Therefore, we propose an aggregated feature difference learning network, AFD-Net, for cross-spectral image patch matching task, as shown in Fig. 2.

Beyond improving feature discrimination, cross-spectral image matching also faces an issue of feature invariance across different domains. Due to the significant appearance changes between cross-spectral images, learning/designing an domain invariant feature is non-trivial, and few studies have addressed this problem. Emerging studies [23, 33] reported that instance normalization (IN) was potential to eliminate the appearance difference. Therefore, we introduce a domain invariant feature extraction network by applying IN which does not only reduce the difference caused by domain changes, but also the illumination variation in single spectral images. In addition, we find the widely used Softmax loss is not the best for our method, because it only encourages the feature separability in Euclidian space but neglects the discrimination [20, 28, 34, 36]. Instead, the matching problem requires separability for larger inter-class distance and also discrimination for smaller intra-class distance. Hence, we borrow the large margin cosine loss (LMCL) [34] in face recognition to optimize AFD-Net. Unlike Softmax loss, LMCL learns features in cosine space, minimizes intra-class distance and maximizes inter-class distance between the matching and non-matching samples.

In short, our contribution in this work has threefold:

(1) We propose an aggregated feature difference learning network, AFD-Net, for cross-spectral image patch matching, in which the feature differences from multiple levels contribute more learning signal to boost up the matching performance.

(2) We introduce a domain invariant feature extraction network by involving Instance Normalization (IN) which can remove the image appearance difference caused by different spectrum and the illumination variation.

(3) Experiments show that our method outperforms the state-of-the-arts on both cross-spectral and single-spectral patch matching benchmarks.

## 2. Related work

### 2.1. Deep learning-based methods

Deep learning-based image matching methods are mainly categorized into two types: *descriptor learning* and *metric learning*. They extract the deep features of image patches through convolutional networks, and then measured the similarity of features by feature distance or metric network [2, 3, 5, 11, 16, 22, 25, 30, 32, 38].

As a pioneer in the descriptor learning, the Siamese network [30] uses two CNN branches with the same structure and shared weights to learn discriminative features for comparing a pair of image patches, and is optimized by the hinge embedding loss. Unlike the pairwise comparison, Balntas *et al*. [5] proposed a PN-Net that adopts triplet comparison to improve the matching performance and the speed of convergence, which is achieved by enforcing the distance of matching pairs must be smaller than any of non-matching pairs using a softPN loss. Later, Aguilera *et al*. [3] directly applied the PN-Net into the cross-spectral image patch matching and proposed a Quadruplet network (Q-Net). Instead, L2-Net [32] and HardNet [22] address the image matching problem from a perspective of mining hard samples. They proposed the exhaustively negative sampling strategy in a mini-batch, and selected the hard negative samples as the major training data. These strategies perform pretty well and reach the state-of-the-art in single spectral image patch matching. Meanwhile, Vijay Kumar *et al*. [16] introduced a global loss into the pairwise and triplet comparison networks, which aims to minimize the mean feature distance of matching samples, maximize the mean distance of non-matching and minimize the variance of intra-class distance.

Alternatively, the metric learning transforms the matching task into a binary classification task by adding a metric network after the feature extraction network. The output is the matching labels. MatchNet [11] is one of the first metric learning methods, which utilizes a Siamese network for feature extraction, and predicts the matching label through

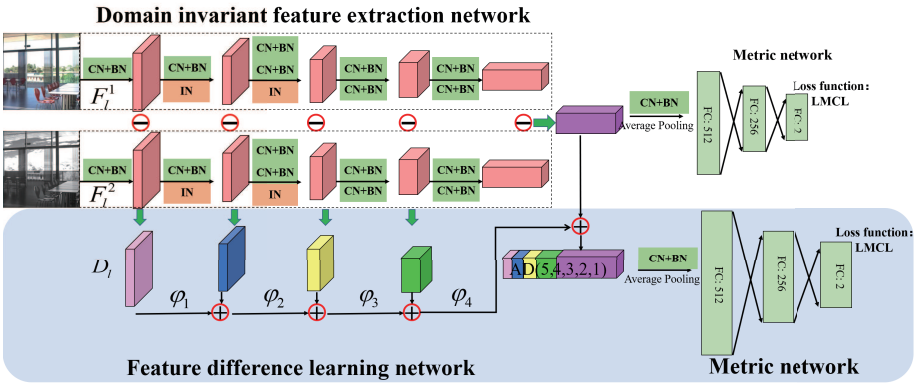



Figure 2. The proposed framework of aggregated feature difference learning network. It has three components: *domain invariant feature extraction network*, *metric network*, and *feature difference learning network*. The domain invariant feature extraction network is for extracting the feature of image patch-pairs by convolution (CN), the metric network is for inferring the matching labels, and the feature difference learning network is for extracting the feature difference at multi-level. In shallow layers of framework, the Instance normalization (IN) and Batch normalization (BN) are used for extracting invariant and discriminative features. The entire framework is jointly optimized by the two large margin cosine loss functions (LMCL).

a fully connected network. Zagoruyko and Komodakis [38] analyzed various network architectures for comparing image patches, *i.e*. Siamese network, Pseudo-Siamese network and 2-channel, and concluded that 2-channel network achieved the best performance.

Whereas, the most of above methods focus on the single spectral image matching. Very few studies considered the problem in cross-spectral domain. Aguilera *et al*. [2] directly applied the Siamese network, Pseudo-Siamese network and 2-channel network for cross-spectral image matching. Later, Quan *et al*. [25] proposed a SCFDM method that learned invariant feature across different domains through a shared feature space.

It is worth noting that all above methods consider only the high-level features, but neglect the effective information of low-level features. By contrast, we found the feature differences in other layers can amplify useful signal to boost up the feature discrimination. Thus, we propose an aggregated feature difference learning network for cross-spectral image patch matching.

## 2.2. Normalization methods

Ioffe and Szegedy [14] introduced the batch normalization (BN) to enable greater learning rate and faster convergence of CNN training by reducing the internal covariate shift. Numerous studies have reported its superiority on many computer vision tasks [18, 26, 39]. Therefore, it has become a default component in many well-known networks, *e.g*. Inception [31], ResNet [12] and DenseNet [13]. Not surprisingly, it is also employed by HardNet for the single spectral image matching [22]. Although, the discriminative features are preserved, the BN-based CNNs are vulnerable to appearance change [23].

Unlike BN, Instance Normalization (IN) is robust to appearance changes, which normalizes the feature by the mean and variance of an instance during both training and test phrases. IN is often applied to the style transfer tasks due to its capability of removing the instance-specific contrast information [33]. Unfortunately, IN interferes the feature discrimination.

In this work, we carefully integrate the IN and BN in feature extraction network to take both of their advantages, *i.e*. being invariant across different domains or illumination changes and preserving sufficient discrimination. Compared with other domain adaption methods, the combination of IN and BN is simple and effective, which can remove the spectral difference without additional computational cost.

## 2.3. Loss functions

Loss function play a critical role in image matching problem, which determines the training speed and performance of network. The well-accepted loss for metric learning is Softmax loss. However, Softmax loss solely emphasizes the separability of features with different labels, and is insufficient to maximize the feature discrimination for classifying the hard samples. Many emerging loss functions have been proposed to decrease the intra-class variance (compactness) and increase the inter-class distance (separable). Wen *et al*. [36] proposed a center loss to make intra-class compact. Liu *et al*. [20] proposed Angular-softmax (A-softmax) loss to learn the angularly discriminative feature by normalizing the weights. Meanwhile, they introduced the angular margin to reinforce the separability of inter-class. Later, Wang *et al*. [34] proposed a large margin cosine loss (LMCL) that upgraded A-softmax by normalizing the weights and feature vectors, and introduced a cosine margin between decision boundaries. In this paper, we adopt LMCL to optimize our network.

## 3. The proposed network

To tackle the cross-spectral image patch matching problem, we propose an aggregated feature difference learning network (AFD-Net), as shown in Fig. 2. AFD-Net is composed of two sub-networks: the upper one has a domain invariant feature extraction network and a metric network, the loss function is large margin cosine loss (LMCL); the lower one is our feature difference learning network, which aggregates multi-level feature differences from upper sub-network for more discriminative information. The details are given in below.

### 3.1. Feature difference learning network

Siamese network [2] is a successful architecture for a wide bank of vision tasks. Therefore, we adopt the Siamese network in this work for feature extraction. A standard Siamese matching network has a two-branch feature extraction network, which share the weights, see in the upper of Fig. 2. Given a pair of image patches $(P^1, P^2)$, the feature extraction network hierarchically extract features $(F_l^1, F_l^2), l = 1, \ldots, L$, using convolutional blocks, which is composed of convolutional layers, normalization layers, and activation functions. Afterward, the metric network infers the matching label $\hat{y}$ according to the high-level features from the last conventional block.

Conventional methods directly compare the feature maps of two image patches by concatenating them along channels. For matching samples, there exists a large feature variance among different samples due to the different patch contents, which results in a large intra-class variance. On the contrary, the feature difference of matching samples could cancel this variance and reduce the intra-class distance. Meanwhile, it also can amplify the difference (useful learning signal) of non-matching patches. Therefore, a more discriminative feature can be obtained by aggregating those feature differences (FDs) from high-level to low-level. Hence, we propose an aggregated feature difference learning network (AFD-Net) for a better performance, please refer to the lower half of Fig. 2. Specifically, we aggregate the difference of feature maps at multiple levels, $AD$. One can see it has richer information for training (see the bottom row in Fig. 1). The network predicts the matching label based on the aggregated feature difference:

$$\hat{y} = M(AD), \tag{1}$$

where $M(\cdot)$ is the metric network.

In order to keep feature invariance but rich discriminative information, we aggregate FDs from high to low. According to the data characteristic, the FD aggregation can be flexible, such as two levels, three levels or more levels,

$$\begin{aligned} &AD(L, L-1) = \varphi_{L-1}(D_{L-1}) \oplus D_L, \\ &AD(L, L-1, L-2) = \varphi_{L-1}(\varphi_{L-2}(D_{L-2}) \oplus D_{L-1}) \oplus D_L, \\ &AD(L, L-1, L-2, ...1) = \varphi_{L-1}(...\varphi_1(D_1) \oplus D_2...) \oplus D_L, \end{aligned} \tag{2}$$

where $D_l$ represents the feature difference in $l_{th}$, $D_l = |F_l^1 - F_l^2|$. $\varphi_l(\cdot)$ denotes the process of $l_{th}$ convolutional block, which can re-extract the feature from feature difference, meanwhile, unify the size of feature maps at two adjacent levels. $\oplus$ is an operation of concatenating two feature maps along channels.

Both upper and lower sub-networks are jointly optimized. In training process, the upper one guides the learning process of feature extraction network, the lower one mainly optimizes the learning of aggregated feature difference. And the output of lower one is the result of whole framework.

### 3.2. Domain invariant feature extraction network

Since cross-spectral images are formed by different imaging mechanisms, the pixels and low-level features of different spectral images preserve the private properties of corresponding domains. They inherently enlarge the domain feature distance between two images. Cross-spectral image matching expects the extracted feature to be invariant across different domains.

Instance normalization (IN) has been reported a capability of eliminating the appearance change. However, it also drops useful content information and impedes the model capability [23]. On the contrary, Batch Normalization (BN) can significantly speed up training and improve the model performance. Hence, it has been a default component in most prevalent CNN architectures. In this work, we carefully integrate IN and BN into the feature network to extract the domain invariant features without degrading feature discrimination.

As our analysis shows that the properties of domain mainly exist in the low-level features, the majority of high-level feature is the abstract information. Therefore, we apply IN after BN into the shallow layers (CN-BN-Relu-IN-Relu) in feature extraction network to reduce the feature variance caused by varied domains, but only BN into deeper layers (CN-BN-Relu) to preserve the feature discrimination. Please refer the domain invariant feature extraction network in Fig. 2 for our setting.

### 3.3. Optimization function

In metric learning methods, the image patch matching task is viewed as a classification problem. Therefore, the widely used loss function is Softmax loss.

Given a training image patch-pair $(P_i^1, P_i^2)$, and its corresponding matching label $y_i$. Based on the training dataset $\{(P_i^1, P_i^2), y_i\}_{i=1...N}$, the network can be optimized by the

Softmax loss function:

$$L_{Softmax} = -\frac{1}{N}\sum_{i=1}^{N}\log p_i,$$
$$p_i = \frac{e^{\|W_{y_i}\|\|h_i\|\cos(\theta_{y_i,i})}}{\sum_{j=1}^{2} e^{\|W_j\|\|h_i\|\cos(\theta_{j,i})}}, \quad (3)$$

where $N$ is the number of training samples, $p_i$ is the posterior probability of the $i_{th}$ training sample that is corresponding to the given label, $h_i$ is the input of the last fully connected layer for the $i_{th}$ samples, $W_j$ is the weights in the $j_{th}$ column of the last fully connected layer, and the corresponding bias is assumed to be zero, $\theta_{j,i}$ is the angle between the $W_j$ and $h_i$.

However, Softmax loss only encourages the feature separability but neglects the discrimination. For matching tasks, it is impractical to pre-collect all the possible samples for training. We expect features could be generalized well for other unseen samples. It requires features to be discriminative enough not just separable. To this end, we adopt the large margin cosine loss (LMCL) to optimize our AFD-Net. LMCL reformulates Softmax loss into cosine space by normalizing the feature vectors and weights using $L_2$ norm, which makes the optimization only depend on the angles and removes radial variance [20, 34]. Moreover, there is a cosine margin $m$ to expend the decision boundary between two categories, which can increase the inter-class separability and decrease the intra-class variation.

$$L_{LMCL} = -\frac{1}{N}\sum_{i=1}^{N}\log(pc_i),$$
$$pc_i = \frac{e^{s\left(\cos(\theta_{y_i,i})-m\right)}}{e^{s\left(\cos(\theta_{y_i,i})-m\right)} + \sum_{j=1,j\neq y_i}^{2} e^{s\cos(\theta_{j,i})}}, \quad (4)$$
$$subject\ to$$
$$W = \frac{W}{\|W\|}, h_i = \frac{h_i}{\|h_i\|}, \cos(\theta_{j,i}) = W_j^T h_i,$$

where the definitions of $W$ and $h$ are similar as Eq. 3, $s$ is the scale parameter, $m$ is the cosine margin.

We applied two LMCLs to jointly optimize the both feature difference network and domain invariant feature extraction network. The optimization is based on the stochastic gradient descent (SGD) and momentum.

# 4. Experiments

To demonstrate the effectiveness of AFD-Net, we evaluate it on a cross-spectral dataset, VIS-NIR patch dataset, and compare it with four handcraft feature methods (SIFT [21], GISIFT [10], EHD [1], LGHD [4]) and eight deep learning state-of-the-arts including Siamese network [2], Pseudo-Siamese network [2], 2-channel network [2], PN-Net [5], Q-Net [3], L2-Net [32], HardNet [22] and SCFD-M [25]. Although AFD-Net is designed for cross-spectral image patch matching, we also test it on a benchmark of single spectral dataset, namely Multi-view stereo correspondence dataset, to illustrate a better generalizability.

| Category | Number | Category | Number | Category | Number |
|---|---|---|---|---|---|
| Country | 277504 | Field | 240896 | Forest | 376832 |
| Indoor | 60672 | Mountain | 151296 | Oldbuilding | 101376 |
| Street | 164608 | Urban | 147712 | Water | 143104 |

Table 1. The number of image patch-pairs of nine categories in cross-spectral image patch matching dataset VIS-NIR.

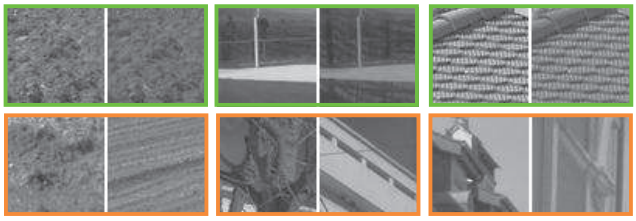


Figure 3. Six image patch-pairs from the cross-spectral dataset. The left is the visible spectrum (VIS)image patches, and the right is the near-infrared (NIR). The first row is matching samples, and the second row is the non-matching samples.

## 4.1. Datasets

*VIS-NIR patch dataset* has been used as a benchmark cross-spectral image patch dataset in [2, 3] for evaluating the metric learning and descriptor learning methods, which were collected from the public VIS-NIR scene dataset by Aguilera [2, 9]. It has nine categories including over 1.6 million VIS-NIR patch-pairs in total, in which each patch has a size of $64 \times 64$. The patches were cropped around the SIFT points in images, the half of VIS image patches and their corresponding NIR image patches form the matching pairs, the other half VIS image patches and the random NIR image patches compose the non-matching pairs. The number of patch-pairs per category is listed in Table 1. Figure 3 shows six samples of patch-pairs from the dataset. Similar to the studies [2,3,25], our framework is also trained on the Country category and test on the remaining categories. It is worth to note that there are significant differences between the categories. Therefore, a satisfied matching performance could be achieved on all test categories when the network has very strong generalization ability.

*Multi-view stereo correspondence dataset* also named as Brown, it is a single spectral image dataset, which consists of corresponding patches sampled from 3D reconstructions [8]. It has three subsets: Liberty, Notredame and Yosemite. Each subset contains 450K, 468k, 634K unique image patches and their corresponding 3D points ID. Each patch was cropped around an interest point, Difference of Gaussian (DOG), with a size of $64 \times 64$. These patches constitute 100K, 200K and 500K labeled pairs, respectively. Half of these pairs are matched, which have the same 3D points ID. The other half are non-matching pairs that have different 3D points ID. The patches in a pair may have no-

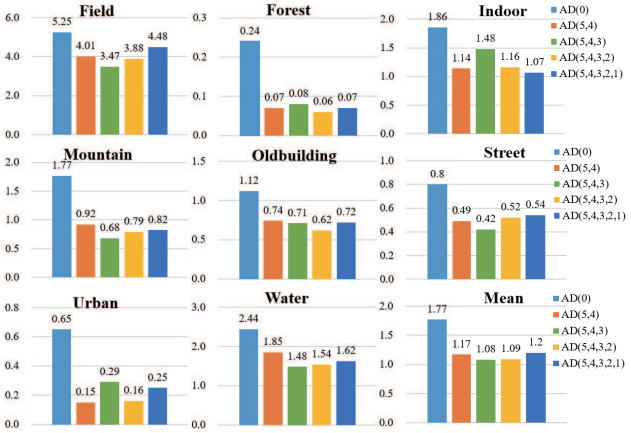



Figure 4. The FPR95 performances of AFD-Net with different aggregation configurations on the VIS-NIR dataset. AD(0) is a general Siamese network, AD($L, L-1, ...$) represents AFD-Net uses the feature difference aggregated from $L_{th}, L-1_{th}, ...$ levels.

table changes in illumination, rotation, translation and perspective. Previous studies [8, 11, 32, 38] have treated it as the standard evaluation dataset. Therefore, we follow these studies to train our framework on one subset and choose 100K samples of the other two subsets for test.

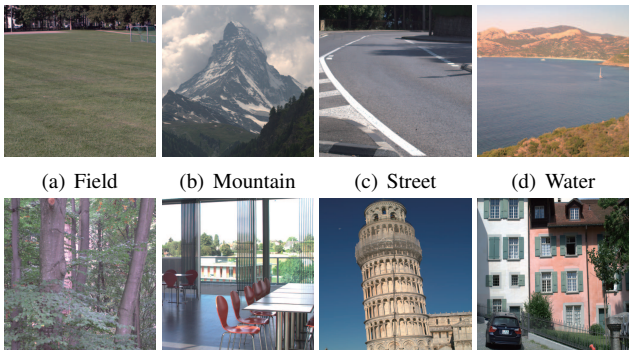
### 4.2. Training

All training and test were implemented on NVIDIA GTX 1080 GPU. The training process is based on the large margin cosine loss (LMCL), which optimized by the stochastic gradient descent (SGD) under the mini-batch. The size of mini-batch is 256, the momentum is 0.9, the initial learning rate is 0.01, with the decay factor 0.9. All samples were normalized to $[0, 1]$, and the data augmentation is carried out through the random flipping, random rotating $(90^{o}, 180^{o}, 270^{o})$ and random contrast change. The false positive rate at 95% recall (FPR95) is employed as evaluation metric of the matching performance [2, 3, 22, 25, 32]. The smaller FPR95 represents the better matching performance.

### 4.3. Ablation study

Since we propose an aggregated feature difference learning network (AFD-Net) with a domain invariant feature extraction network (instance normalization and batch normalization, IBN), and LMCL loss, it is worth to evaluate the effectiveness of these components in the framework. The evaluation is carried on VIS-NIR patch dataset in terms of the FPR95 results and their means on eight test categories.

*AFD-Net*: Since the aggregation of FD can be flexible, we first evaluate our AFD-Net with varied configurations of aggregation. Note that the domain invariant feature network and LMCL are applied in this test. All possible



(a) Field (b) Mountain (c) Street (d) Water

(e) Forest (f) Indoor (g) Oldbuilding (h) Urban

Figure 5. VIS images of eight test categories. The images in the first row have less edge and texture features than the second row.

configurations are AD(0), AD(5,4), AD(5,4,3), AD(5,4,3,2) and AD(5,4,3,2,1), in which, AD(0) represents a general Siamese network without any aggregation; AD(5,4) denotes the FDs in the $5_{th}$ and $4_{th}$ levels are aggregated. It is similar for other configurations. The results are shown in Fig. 4. One can see that the performance of AD(0) (Siamese network) is the worst. By contrast, AD(5,4), AD(5,4,3), AD(5,4,3,2) and AD(5,4,3,2,1) all achieve significant improvements. Specifically, the maximal improvements of FPR95 in eight categories are 33.90%, 75.00%, 42.47%, 61.58%, 44.64%, 47.50%, 76.92%, 39.34%, respectively, and the mean improvements raise up from 32.20% to 38.98% by AFD-Net. This demonstrates the better discrimination of aggregated feature difference and the effectiveness on cross-spectral image patch matching task.

Another observation in Fig. 4 is that there exists a general trend of the change of FPR95 when more FDs are aggregated, where the mean of FPR95 decreases initially and then increases. We think the decrease is contributed by the features in middle levels, because they have more texture signal compared with high-level features and are invariant to the spectrum domain, rotation and illuminance. The aggregated FDs further boosts up the feature discrimination. However, the increase of FPR95 is due to low-level feature maps have more texture information but sensitive to the changes at pixel level and less invariant. So, there exists a tradeoff. Specifically, AD(5,4,3) reaches the best matching performance on Field, Mountain, Street and Water categories. Instead, AD(5,4,3,2) performs better on Forest and Oldbuilding and almost the best on Urban. AD(5,4,3,2,1) has the best result on Indoor category. We think these categories have much more edge and texture information, as shown in Fig. 5, which leads AD(5,4,3,2) and AD(5,4,3,2,1) to have an improved performance. As AD(5,4,3) is the best on average, we adopt this configuration to AFD-Net in subsequent experiments.

| Settings | Field | Forest | Indoor | Mountain | Oldbuilding | Street | Urban | Water | Mean |
|---|---|---|---|---|---|---|---|---|---|
| NO-LMCL | 6.17 | 0.20 | 2.13 | 2.86 | 1.15 | 0.85 | 0.75 | 2.58 | 2.09 |
| BN-LMCL | 4.36 | 0.09 | 2.04 | 1.63 | 0.76 | 0.62 | **0.28** | 2.01 | 1.47 |
| IBN-LMCL | **3.47** | **0.08** | **1.48** | **0.68** | **0.71** | **0.42** | 0.29 | **1.48** | **1.08** |
| IBN-Softmax | 4.43 | 0.09 | 2.78 | 1.17 | 1.65 | 0.66 | 1.53 | 2.24 | 1.82 |

Table 2. The FPR95 performances of AFD-Net on VIS-NIR dataset with different normalization methods and loss functions. The normalization methods includes: no normalization "NO", only with batch normalization "BN" and combining instance normalization (IN) and BN "IBN". The loss functions includes the Softmax loss "Softmax" and large margin cosine loss "LMCL". The best performance is in bold.

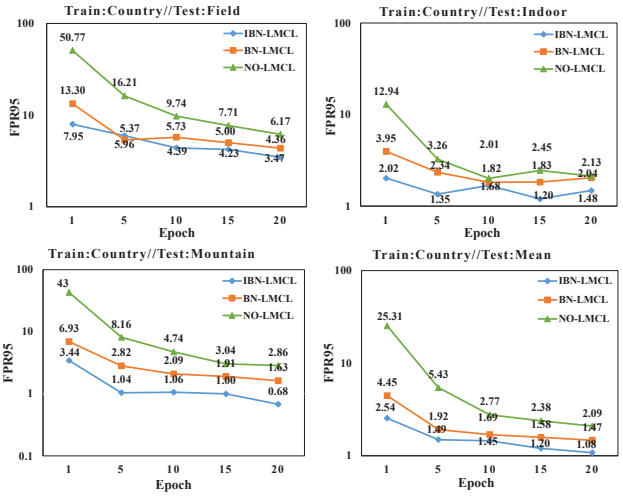



Figure 6. The training efficiency of AFD-Net using different normalizations. IBN achieves a faster convergence and the best FPR95.

*Normalization*: As our domain invariant feature extraction network uses Instance Normalization (IN) and Batch Normalization (BN) to eliminate the domain variance and preserve the discriminative information, we set up three configurations that all use large margin cosine loss (LMCL). They are no any normalization, NO-LMCL; using Batch Normalization only, BN-LMCL; and using both Instance and Batch Normalizations, IBN-LMCL. The comparison result is listed in the Table 2. One can see that IBN-LMCL performs the best on seven categories except for Urban, but is in the second place with a very small margin. The mean of FPR95 shows that IBN-LMCL improves the accuracy 48.33% and 26.53% up than NO-LMCL and BN-LMCL, respectively. This result confirms that IN does eliminate certain domain properties, and BN preserves discriminative information. We also tested the training efficiency using different normalizations, and plot FPR95 against training epoch in Fig. 6. It clearly shows that the domain invariant feature extraction network (IBN-LMCL) achieves faster training convergence.

*Loss function*: As the loss function determines the ultimate goals of network learning, we validate the effectiveness of LMCL by comparing with Softmax loss. There are two parameters, $(s, m)$, in LMCL loss, we first evaluate them according to the principles in the previous study [34]. The result in Table 3 shows the optimal parameters are "$s = 20, m = 0.25$". Thus, we will use this setting in subsequent experiments. Comparing with the results using Softmax loss (see Table 2), IBN-LMCL outperforms IBN-Softmax on all categories, especially on Indoor, Old-building and Urban, their matching accuracies raise up 46.76%, 56.97% and 81.05%, respectively. And the average FPR95 is decreased 40.66% using IBN-LMCL. This result empirically show that LMCL is more suitable for matching problems than Softmax loss.

| $s$ | $m$ | Field | $s$ | $m$ | Field | $s$ | $m$ | Field |
|---|---|---|---|---|---|---|---|---|
| 10 | 0.15 | 3.66 | 10 | 0.25 | 4.20 | 10 | 0.35 | 4.05 |
| 20 | 0.15 | 4.12 | **20** | **0.25** | **3.47** | 20 | 0.35 | 4.86 |
| 30 | 0.15 | 3.99 | 30 | 0.25 | 4.21 | 30 | 0.35 | 4.10 |

Table 3. The FPR95 when varying parameters $(s, m)$ of LMCL. AFD-Net was trained on *Country* category and tested on *Field* category.

## 4.4. Cross-spectral image matching

To demonstrate the effectiveness of AFD-Net on cross-spectral image patch matching problem, we compare it with twelve state-of-the-arts, and list the results in Table 4. One can see AFD-Net outperforms other methods on all test categories. Specifically, it improves matching performance 61.43% up in term of the mean FPR95 than HardNet [22].

It is worth to note that HardNet [22] and SCFDM [25] are in the second and third places, respectively. SCFDM is particularly designed for cross-spectral image matching by learning the feature from a shared feature space through the spatial connected mode and a feature discrimination constrain. HardNet applies a exhaustive hard sample mining for training, which enforces the network to learn more discriminative features. However, they only utilize the high-level features. Meanwhile, HardNet is lack of the feature invariance across different spectrums, and its loss function just emphasized the local margin between the matching samples and non-matching samples. Analogically, the loss function used in SCFDM cannot minimize the intra-class distance either. Compared with them, AFD-Net aggregates feature differences from multiple levels to amplify the useful learning signal, removes the spectral difference by domain invariant normalization, and make intra-class distance more compact by LMCL. Therefore, AFD-Net outperforms them on cross-spectral image patch matching task.

| Models | Field | Forest | Indoor | Mountain | Oldbuilding | Street | Urban | Water | Mean |
|---|---|---|---|---|---|---|---|---|---|
| Traditional methods | | | | | | | | | |
| SIFT [21] | 39.44 | 11.39 | 10.13 | 28.63 | 19.69 | 31.14 | 10.85 | 40.33 | 23.95 |
| GISIFT [10] | 34.75 | 16.63 | 10.63 | 19.52 | 12.54 | 21.80 | 7.21 | 25.78 | 18.60 |
| EHD [1] | 33.85 | 19.61 | 24.23 | 26.32 | 17.11 | 22.31 | 3.77 | 19.80 | 20.87 |
| LGHD [4] | 16.52 | 3.78 | 7.91 | 10.66 | 7.91 | 6.55 | 7.21 | 12.76 | 9.16 |
| Descriptor learning | | | | | | | | | |
| PN-Net DA [5] | 20.09 | 3.27 | 6.36 | 11.53 | 5.19 | 5.62 | 3.31 | 10.72 | 8.26 |
| Q-Net DA [3] | 17.01 | 2.70 | 6.16 | 9.61 | 4.61 | 3.99 | 2.83 | 8.44 | 6.91 |
| L2-Net DA [32] | 16.77 | 0.76 | 2.07 | 5.98 | 1.89 | 2.83 | 0.62 | 11.11 | 5.25 |
| HardNet DA [22] | 10.89 | 0.22 | 1.87 | 3.09 | 1.32 | 1.30 | 1.19 | 2.54 | 2.80 |
| Metric learning | | | | | | | | | |
| Siamese DA [2] | 15.79 | 10.76 | 11.60 | 11.15 | 5.27 | 7.51 | 4.60 | 10.21 | 9.61 |
| Pseudo-Siamese DA [2] | 17.01 | 9.82 | 11.17 | 11.86 | 6.75 | 8.25 | 5.65 | 12.04 | 10.31 |
| 2-channel DA [2] | 9.96 | 0.12 | 4.40 | 8.89 | 2.30 | 2.18 | 1.58 | 6.40 | 4.47 |
| SCFDM DA [25] | 7.91 | 0.87 | 3.93 | 5.07 | 2.27 | 2.22 | 0.85 | 4.75 | 3.48 |
| AFD-Net DA | **3.47** | **0.08** | **1.48** | **0.68** | **0.71** | **0.42** | **0.29** | **1.48** | **1.08** |

Table 4. The comparison of FPR95 among our proposal and twelve state-of-the-art methods on VIS-NIR scene dataset. All methods were trained on country category and tested on the other eight categories. DA denotes using the data augmentation in training process. The best performance is in bold.

| Training | Notredame | Yosemite | Liberty | Yosemite | Liberty | Notredame | |
|---|---|---|---|---|---|---|---|
| Test | Liberty | | Notredame | | Yosemite | | Mean |
| TNet-TGLoss DA [16] | 9.91 | 13.45 | 3.91 | 5.43 | 10.65 | 9.47 | 8.80 |
| TNet-TLoss DA [16] | 10.77 | 13.90 | 4.47 | 5.58 | 11.82 | 10.96 | 9.58 |
| SNet-Gloss DA [16] | 6.39 | 8.43 | 1.84 | 2.83 | 6.61 | 5.57 | 5.27 |
| PN-Net [5] | 8.13 | 9.65 | 3.71 | 4.23 | 8.99 | 7.21 | 6.98 |
| Q-Net DA [3] | 7.64 | 10.22 | 4.07 | 3.76 | 9.34 | 7.69 | 7.12 |
| DeepDesc [30] | 10.90 | | 4.40 | | 5.69 | | 6.99 |
| L2-Net DA [32] | 2.36 | 4.70 | 0.72 | 1.29 | 2.57 | **1.71** | 2.22 |
| HardNet DA [22] | 1.49 | 2.51 | 0.53 | 0.78 | 1.96 | 1.84 | 1.51 |
| MatchNet [11] | 6.90 | 10.77 | 3.87 | 5.67 | 10.88 | 8.39 | 7.44 |
| DeepCompare 2ch-2stream DA [38] | 4.85 | 7.20 | 1.90 | 2.11 | 5.00 | 4.10 | 4.19 |
| DeepCompare 2ch-deep DA [38] | 4.55 | 7.40 | 2.01 | 2.52 | 4.75 | 4.38 | 4.26 |
| SCFDM DA [25] | **1.47** | 4.54 | 1.29 | 1.96 | 2.91 | 5.20 | 2.89 |
| AFD-Net DA | 1.53 | **2.31** | **0.47** | **0.72** | **1.63** | 1.88 | **1.42** |

Table 5. The comparison of FPR95 among our proposal and twelve state-of-the-art methods on Multi-view stereo correspondence dataset.

### 4.5. Multi-view stereo matching

To demonstrate the generalizability of our proposal, we also compare AFD-Net with twelve state-of-the-art methods on a single spectral image dataset, *i.e*. multi-view stereo correspondence dataset [37]. Results are listed in Table 5.

One can see AFD-Net outperforms the other methods on average again. Especially, it performs the best when the training dataset is Liberty and Yosemite. Thanks to exhaustively hard sampling strategy, HardNet and L2-Net also achieve rather good matching performances, and are in the second place and third place, respectively. AFD-Net reduces the average FPR95 compared with HardNet and L2-Net by 5.96% (from 1.51 to 1.42) and 36.04% (from 2.22 to 1.42). It shows the performance improvement between AFD-Net and HardNet on the single-spectral dataset is less than on the cross-spectral dataset. We believe that the single-spectral images have no domain difference, thus, our domain invariant feature extraction network contribute less for patch matching task. However, our aggregated feature difference and LMCL loss still let AFD-Net outperform HardNet and L2-Net without using hard sampling strategy. This result also demonstrate a better generalizability of AFD-Net.

## 5. Conclusion

We propose an aggregated feature learning network (AFD-Net), which utilizes the multi-level feature difference and learns more useful signal from FDs for cross-spectral image patch matching task. In addition, we introduce a domain invariant feature extraction network using instance normalization (IN) and batch normalization (BN). IN can remove the spectral changes in the cross-spectral images and the illumination changes in single spectral images, and the BN can preserve the discriminative features. To further enhance the feature discrimination, we borrow the large margin cosine loss (LMCL) for network optimization. Evaluation experiments were conducted on both the cross-spectral image patch matching dataset (VIS-NIR) and the singe spectral image patch matching dataset. The results demonstrate that AFD-Net achieves the state-of-the-art matching performance. In the future work, we are going to investigate a complete and efficient methodology of hard sample mining for AFD-Net.

## References


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

[2] Cristhian A. Aguilera, Francisco J. Aguilera, Angel D. Sappa, Cristhian Aguilera, and Ricardo Toledo. Learning cross-spectral similarity measures with deep convolutional neural networks. In *CVPR*, pages 1–9, 2016.

[3] Cristhian A. Aguilera, Angel D. Sappa, Cristhian Aguilera, and Ricardo Toledo. Cross-spectral local descriptors via quadruplet network. *Sensors*, 17(4):873, 2017.

[4] Cristhian A. Aguilera, Angel D. Sappa, and Ricardo Toledo. Lghd: A feature descriptor for matching across non-linear intensity variations. In *IEEE ICIP*, 2015.

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

[6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer Vision Image Understanding*, 110(3):346–359, 2008.

[7] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24(4):509–522, 2002.

[8] Matthew Brown, Gang Hua, and Simon Winder. Discriminative learning of local image descriptors. *IEEE TPAMI*, 33(1):43–57, 2011.

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

[10] Damien Firmenichy, Matthew Brown, and Sabine Ssstrunk. Multispectral interest points for rgb-nir image registration. In *ICIP*, pages 181–184, 2011.

[11] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, pages 3279–3286, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645, 2016.

[13] Gao Huang, Zhuang Liu, Laurens Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017.

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. *ICML*, pages 448–456, 2015.

[15] Felix Juefei-xu, Dipan K. Pal, and Marios Savvides. Nir-vis heterogeneous face recognition via cross-spectral joint dictionary learning and reconstruction. In *CVPR*, pages 141–150, 2015.

[16] B. G. Vijay Kumar, Gustavo Carneiro, and Ian Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *CVPR*, pages 5385–5394, 2016.

[17] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[18] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2016.

[19] Peizhong Liu, Jingming Guo, Chiyi Wu, and Danlin Cai. Fusion of deep learning and compressed domain features for content based image retrieval. *IEEE TIP*, 26(12):5706–5717, 2017.

[20] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 6738–6746, 2017.

[21] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[22] Anastasiya Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *preprint arXiv:1705.10872*, 2017.

[23] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, pages 464–479, 2018.

[24] Peter Pinggera, Toby Breckon, and Horst Bischof. On cross-spectral stereo matching using dense gradient features. In *CVPR*, 2012.

[25] 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 *ACCV*, 2018.

[26] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE TPAMI*, 2017.

[27] Jrgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.

[29] Steven M. Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *CVPR*, pages 519–528, 2006.

[30] Edgar Simoserra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Morenonoguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, pages 118–126, 2015.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[32] Yurun Tian, Bin Fan, and Fuchao Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *CVPR*, 2017.

[33] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, pages 4105–4113, 2017.

[34] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, , and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. *preprintarXiv:1801.09414*.

[35] Shuang Wang, Dou Quan, Xuefeng Liang, Mengdan Ning, Yanhe Guo, and Licheng Jiao. A deep learning framework for remote sensing image registration. *ISPRS Journal of Photogrammetry Remote Sensing*, 2018.

[36] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515, 2016.

[37] Simon Winder, Gang Hua, and Matthew Brown. Picking the best daisy. In *CVPR*, pages 178–185, 2009.

[38] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, pages 4353–4361, 2015.

[39] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *CVPR*, 2018.