Information Fusion 91 (2023) 623–632

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Spectral-invariant matching network

Yeongmin Ko [a], Yong-Jun Jang [b], Vinh Quang Dinh [c], Hae-Gon Jeon [d], Moongu Jeon [a,*]

[a] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, 61005, Republic of Korea
[b] Hyundai Rotem, 37, Cheoldobangmulgwan-ro, Uiwang-si, Gyeonggi-do, 16082, Republic of Korea
[c] School of Electrical Engineering and Computer Science, Vietnamese-German University, Thu Dau Mot 75000, Viet Nam
[d] Artificial Intelligence Graduate School and the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, 61005, Republic of Korea

ARTICLE INFO

Keywords:
Sensor fusion
VIS–NIR
Image matching
Stereo matching
Domain conversion

ABSTRACT

As the need for sensor fusion systems has grown, developing methods to find correspondences between images with different spectral ranges has become increasingly important. Since most images do not share low-level information, such as textures and edges, existing matching approaches fail even with convolutional neural networks (CNNs). In this paper, we propose an end-to-end metric learning method, called SPIMNet (SPectral-Invariant Matching Network) for robust cross- and multi-spectral image patch matching. While existing methods based on CNNs learn matching features directly from cross- and multi-spectral image patches, SPIMNet transforms across spectral bands and discriminates for similarity in three steps. First, (1) SPIMNet is adjusted for a feature domain by introducing a domain translation network; then (2) two Siamese networks learn to match the adjusted features with the same spectral domain; and (3) the matching features are fed to fully-connected layers to determine the identity of the patches as a classification task. By effectively incorporating each step, SPIMNet achieved competitive results on a variety of challenging datasets, including both VIS–NIR and VIS–Thermal image pairs. Our code is available at https://github.com/koyeongmin/SPIMNet.

1. Introduction

Many researchers and industries utilize sensors of various domains to get more information about targets. For managing information from each sensor, proper sensor fusion methods are required. In the field of computer vision, cross-spectral (*i.e.*visible–near infrared (NIR)) and multi-spectral (*i.e.*visible–thermal) image matching are being actively studied because the different spectral domains can provide complementary information [1]. As an example, visible and thermal images can mutually compensate for rich color information and high textural structures in low-light conditions, making these images suitable for all-day vision systems [2]. All-day vision or fusing multi-spectral images has become an essential and significant task for sensor fusion systems that conduct facial expression recognition [3,4], material classification [5,6], medical image analysis [7], pansharpening [8], and pedestrian detection [9–11].

Since cross- and multi-spectral images capture different wavelength spectral ranges, the images appear significantly different in both intensity and pixel levels. Even with well-known local feature descriptors [12,13], the relationship between images across spectral domains cannot be accounted for, which results in severe performance drops in matching tasks. Recently, convolutional neural networks (CNNs) have demonstrated some ability to address this issue, by leveraging semantic information along with low-level features. Siamese structures overcome the somewhat challenging matching problems among various spectral domains [14–17]. In most siamese structures, the same deep neural network is applied to both multi-spectral image patches and extracts each feature. Their loss functions make the distance between two positive patches short, otherwise far. Encoder–decoder structures are also utilized to extract common features between multi-spectral image patches [18,19].

Although these previous methods show that their methods work for fusing cross- and multi-spectral images, these methods separately extract features from each image, and just aggregate the features to fuse information or to predict similarities. We have observed that the methods are not suited to dealing with both intensity- and pixel-level differences because these differences can be one reason to make a fusion between multi-spectral images hard and previous methods do not have any module to reduce the difference. In this paper, we present a SPectral-Invariant Matching Network (SPIMNet), an end-to-end CNN

* Corresponding author.
E-mail addresses: koyeongmin@gist.ac.kr (Y. Ko), yjchang@hyundai-rotem.co.kr (Y. Jang), vinh.dq2@vgu.edu.vn (V.Q. Dinh), haegonj@gist.ac.kr (H. Jeon), mgjeon@gist.ac.kr (M. Jeon).

https://doi.org/10.1016/j.inffus.2022.10.033
Received 10 February 2022; Received in revised form 29 October 2022; Accepted 31 October 2022
Available online 5 November 2022

framework for robust image patch matching across different spectral domains. These are the primary contributions of this study:

- In contrast to previous methods that extract features directly from input patches, SPIMNet learns the spectral translations of input patches using a domain conversion network. We can use similar features to compare image patches across different spectral domains.
- SPIMNet utilizes a dual-Siamese network for feature extraction from each translated piece of information to predict the matching label through a fully connected network.
- The proposed end-to-end method can be trained from scratch without any pre-trained backbone network, and we obtain competitive results over several standard datasets, including both visible–NIR and visible–thermal images.
- Additionally, ablation studies indicate that each of these technical contributions leads to appreciable improvements in matching accuracy, and we show that the proposed method can be applied to various applications such as stereo matching and image enhancement.

2. Related works

Hand-crafted Feature Descriptions hand-crafted features such as SIFT [12], SURF [13] and FAST [20] are based on measurements of texture similarities and have shown promise for finding correspondences between visible images, even with illumination and scale changes. A modification of the hand-crafted features was used to handle the issue of dense correspondences in [21]. However, these methods often fail in cross- and multi-spectral imagery because their different spectral characteristics result in nonlinear variations in intensity, and inconsistent textures.

Alternative methods, such as the multi-spectral SIFT [22] improved the performance of scene category recognition for a pair of VIS–NIR scenes by analyzing the statistical dependencies between them. In [23], image features between visible and thermal images were extracted using a frequency-based detector and described using a combination of spatial and frequency information. For dense matching, modified hand-crafted matching costs were designed. In [24], a selective normalized cross-correlation established dense pixel correspondences in the input multi-spectral images, and its mathematical parameterization was proposed to make the optimization tractable. Heo et al. [25] analyzed a color intensity model and proposed an adaptive normalized cross-correlation for stereo matching, and their extended work [26] iteratively estimated dense depth maps and adjusted color consistency. In [27], cross-spectral stereo matching was presented with dense gradient features based on the HOG descriptor [28]. Kim et al. [29] proposed a dense descriptor for cross-spectral correspondences with their adaptive self-correlation and randomized receptive field pooling. Holloway et al. [30] presented an assorted camera array and a normalized gradient cost to measure correspondences in cross-spectral images. Kim et al. [31] proposed the dictionary learning method from small patches of the whole-size image for multimodal image fusion. However, hand-crafted features should be re-designed with every characteristic of sensors, and it consumes much effort [32].

CNN-based Multi-spectral Image Matching CNNs have achieved great success in image patch matching, thereby significantly improving state-of-the-art stereo matching [33] and similarity computation [34]. In particular, Siamese structures have demonstrated robustness when performing image matching for various datasets in [35]. A generalization of the Siamese structure in [36] showed promising performance over hand-crafted features by simultaneously learning local patch representation and performing feature comparisons.

Aguilera et al. [16] shows some analysis of three CNN models (2ch, siamese, pseudo-siamese), and these structures are utilized as a typical baseline in follow-up studies. Quan et al. [15] measured the similarity of multi-spectral image patches with shared semantic features. They also introduced AFD-Net, which learns an aggregation of the multi-level feature differences to enhance a discrimination [14]. Beaupre et al. [37] extracted features from each patch using two CNNs that have no weight sharing and applied two fusion operations such as correlation and concatenation. Yu et al. [38] proposed the multi-branch feature extraction network. Because this method can extract various features from the multi-branch network, the method gets the same effect as an ensemble.

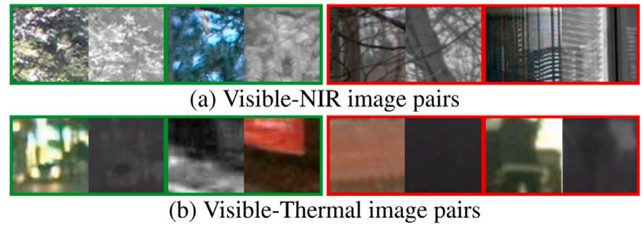


Fig. 1. Examples of cross-spectral image patches. The green and red boxes represent positive and negative samples, respectively. The cross-spectral patches sometimes have visually different appearances in the positive samples and are semantically the same in negative samples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Unsupervised methods can resolve that labeling is sometimes hard when multiple sensors are utilized. Yan et al. [39] used the Expectation Maximization(EM) optimization method to optimize five non-differentiable properties for correspondence matching. Ye et al. [40] presented the unsupervised method to predict transformation parameters between multi-spectral images for registration. The method can generate much data using augmentation, and the multiscale framework achieves robustness to some geometric distortions.

Although multi-spectral images have a different appearance, most previous CNN-based methods did not provide any module and loss functions to dealing the difference. They directly compare the features of each multi-spectral image. It relies solely on CNN's ability to extract features. In our proposed method, we resolve this problem using the domain conversion module. We can get images converted to other domains from the module, and the network can predict results using similar domain images.

Domain conversion Generative Adversarial Networks (GAN) is one of the famous methods for domain conversion [7,41–44]. In most GAN-based methods, the generator gets both the source image and target domain image and generates fused images and the discriminator judges the quality. Because these methods generated directed completely fused images, however, we should divide network architecture and training steps into several streams if we want to utilize low-level features between multi-spectral images or apply them to other applications. Zhi et al. [45] presented a weakly supervised learning framework for dense depth computation from visible and NIR image pairs. We note that the work in [45] also adopted a spectral domain transfer to make pseudo-visible images from NIR and vice versa for only left–right consistency checks in stereo matching. Compared to [45], the domain conversion in SPIMNet learns to make a translated feature prior to encoding the image feature in the end-to-end training step for the image patch matching task. In Section 4, we will demonstrate the effectiveness of this domain conversion for multi-spectral patch matching.

3. Spectral-invariant matching network

Previous works [14,15] on cross-spectral matching have directly extracted discriminative features from input image patches. As shown in Fig. 1, matching image patches from different spectral domains is a challenging task because the objects and materials have totally different

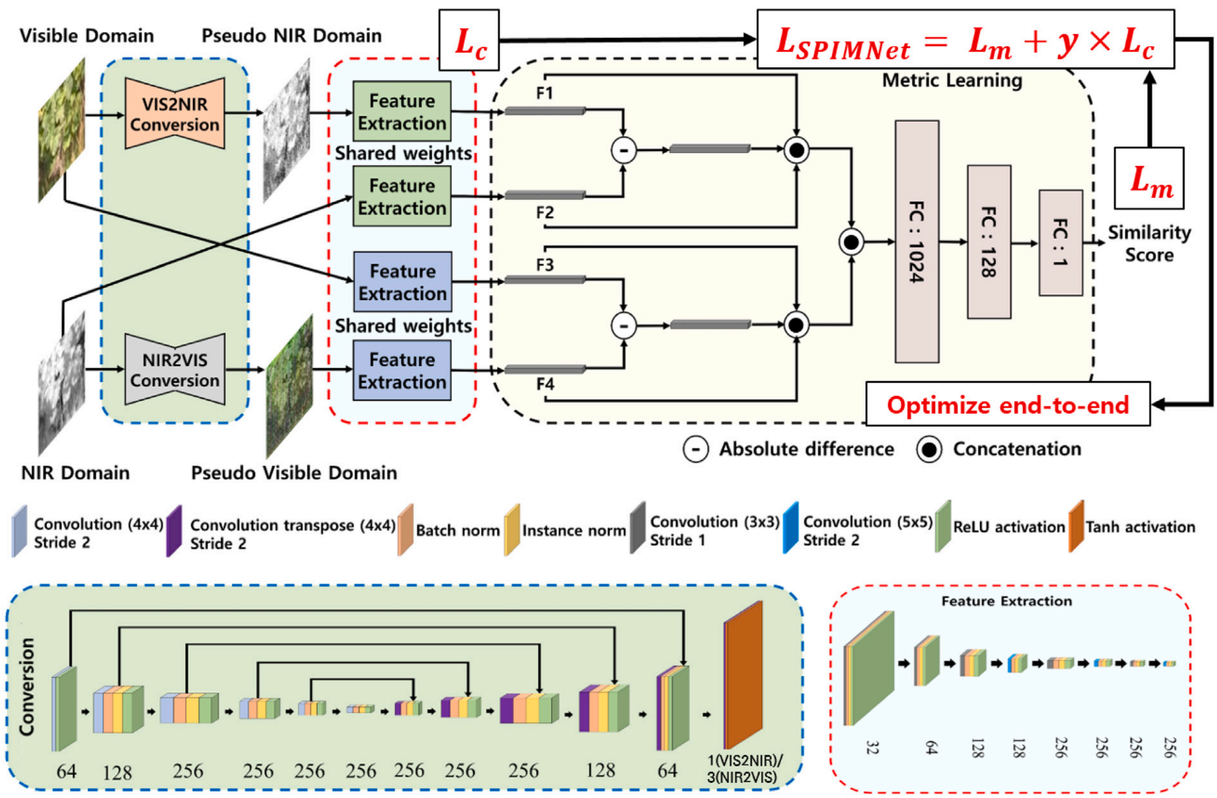


Fig. 2. An overview of SPIMNet and details of its sub-networks. SPIMNet consists of two domain conversion networks (VIS2NIR and NIR2VIS for cross-spectral image matching), two feature extraction networks, and a metric learning network. The feature extraction networks extract discriminative features from an input image and a converted image from the domain conversion network.

appearances. For this reason, performance has been limited in previous works [14,15].

In this work, instead of learning discriminative features directly from cross-spectral image patches, we solve the matching problem with our proposed SPIMNet, which consists of three modules: domain conversion, feature extraction and a metric learning network. An overview of SPIMNet is illustrated in Fig. 2. Note that, for the sake of simplicity, we use two specific domains (visible and NIR) in this section. We will demonstrate that SPIMNet also works well for different types of multi-spectral image pairs, such as visible and thermal, without any modifications in Section 4.

3.1. Network design

Domain Conversion Network The first module is a domain conversion network that translates input images from one domain to another domain, and vice versa. For example, if the input images are a pair of visible and NIR images, the two domain conversion networks make two different translated images. We observe that these translated images play a key role in significantly improving performance across various cross- and multi-spectral datasets.

The domain conversion network is based on U-Net [46] with ten blocks. The encoding blocks consist of convolution, batch norm [47], instance norm [48], and ReLU layers. Although the decoding blocks are similar to the encoding blocks, they use convolution transpose layers instead of convolution layers. The number of filters for the convolution layers in the five encoding blocks and convolution transpose layers in the four decoding blocks are (64, 128, 256, 256, 256, 256) and (256, 256, 256, 128, 64), respectively. All of the convolution and convolution transpose layers use a 4 × 4 filter size with stride 2 and the same padding. The conversion networks are similar except for the last blocks, which depend on the properties of the target domain. For each VIS2NIR and NIR2VIS conversion, the network generates one and three channel outputs, respectively. Tanh activations are used in the VIS2NIR and NIR2VIS conversion networks to normalize a range of the images with [−1, 1].

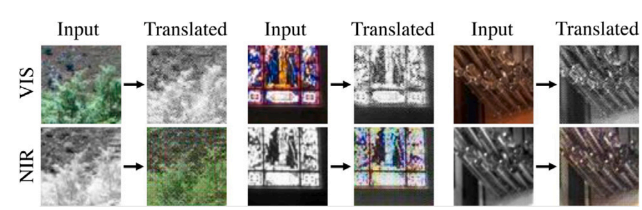


Fig. 3. Qualitative results of the VIS2NIR and NIR2VIS conversion network using the VIS–NIR patch dataset. The left column means input images, and the right columns are translated images.

Fig. 3 shows examples of VIS2NIR and NIR2VIS using the VIS–NIR patch dataset [16]; details are described in Section 4.1. The translated images not only keep the low-level features, but also are similar in appearance to their corresponding input images. Although the intensity levels between the input visible images and the translated images are different, the appearances generated from the domain conversion network alleviate the complex problems encountered in cross-spectral matching.

Feature Extraction Network The second module is a dual-Siamese network for extracting discriminative features from the translated images. Here, each Siamese network learns to extract features coming from the original and the translated domains. The outputs of the two feature extraction networks are four feature vectors, F_1, F_2, F_3, and F_4, which are fed to a metric learning network.

The feature extraction network is comprised of eight layers, each of which includes a convolution, batch norm, instance norm, and

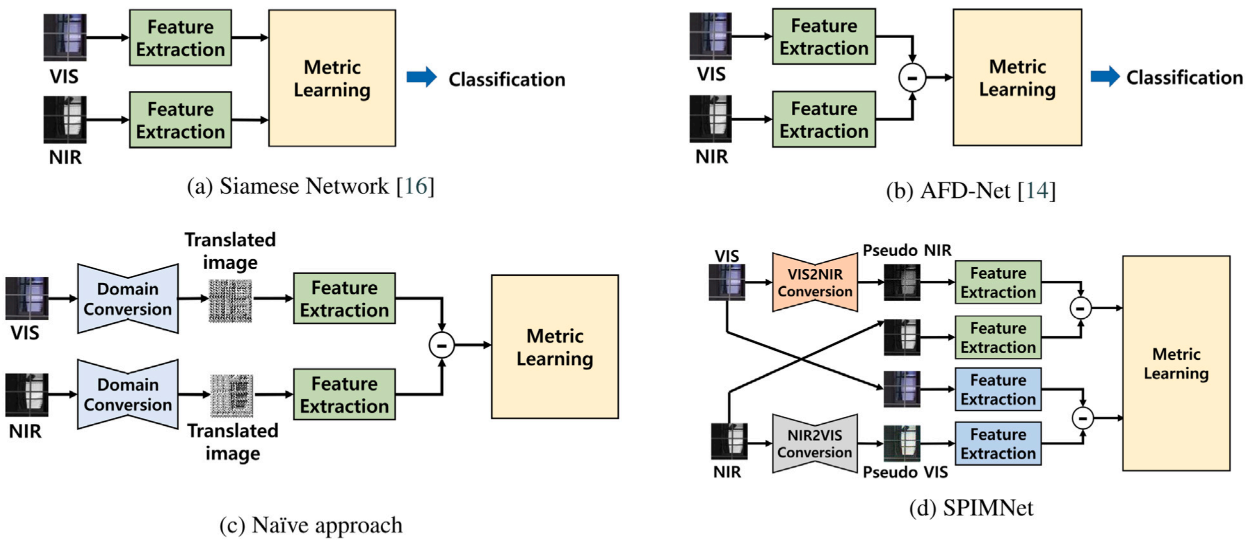


Fig. 4. Network structures for cross-spectral image matching. (a) and (b) Existing approaches learn to extract discriminative features directly from input images. (c) A naïve approach that translates images prior to the extraction of discriminative features. (d) Our proposed approach first translates image patches from one domain to another and vice versa for domain adjustment; after that, its discriminative features can be extracted to compute their similarity.

ReLU activation. The number of filters, filter size, and stride for the convolution layers of the eight blocks are (32, 3 × 3, 1), (64, 3 × 3, 1), (128, 3 × 3, 1), (128, 5 × 5, 2), (256, 3 × 3, 1), (256, 5 × 5, 2), (256, 3 × 3, 1) and (256, 5 × 5, 2), respectively.

Metric Learning Network Finally, the extracted features are fed into a metric learning network which includes fully connected layers. We concatenate several features to fuse rich information. For example, in the NIR domain, two raw features of each input(real NIR and pseudo-NIR) and the difference between the two raw features are concatenated. The concatenate operator can preserve information about each feature. The subtract operator is also frequently utilized by most siamese-based methods because the operator provides useful information in matching tasks that predict whether input patches are in the same location or not [14,38].

In contrast to previous methods, the concatenated features are from similar domains because we utilize the domain conversion network. It can help to improve performance because fusion between similar domains is easier than significantly different domains. The final network output is fed to a nonlinear sigmoid activation function to produce a similarity score between the learned features. SPIMNet makes a binary decision about whether the input pair is associated or not.

3.2. SPIMNet structure rationale

To explain how SPIMNet acts as a robust patch matching network across cross-spectral domains, we must first examine the classic Siamese structure. Fig. 4(a) shows an existing Siamese network in [16]. The network learns to determine the similarities between two patches. AFD-Net [14] in Fig. 4(b) improves the performance of the Siamese network using an advanced manner learning feature, with differences on multiple levels. However, these works are based on features directly extracted from input images when making decisions about whether the input patches are matched. This is not sufficient for CNNs, which need to accurately represent feature maps of the cross-spectral images.

In a naïve manner, we add a domain conversion network to translate input patches with different spectral properties into a common domain, prior to discriminative feature learning in Fig. 4(c). Although the domain conversion network generates matchable patches, one common domain is limited to covering image patches with various low-level and semantic information.

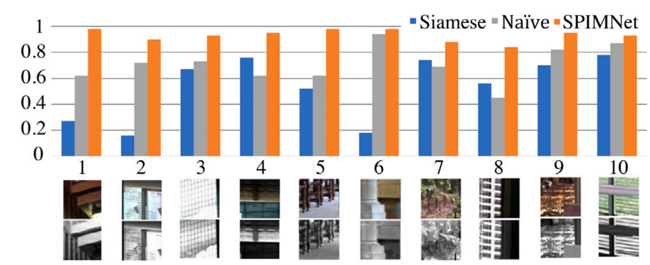


Fig. 5. Similarity scores of Siamese network, naïve approach, and SPIMNet for the 10 hard positive samples.

In cross-spectral matching, each domain has its own characteristics, and their usefulness varies depending on the image contents. To counteract images with different spectral domains, we propose a dual-stream network consisting of two domain conversion networks and four feature learning networks in Fig. 4(d). Since the dual-stream structure allows the network to select the matching domain automatically, SPIMNet can leverage its performance with various images.

The examples in Fig. 5 explain how this is true. We first pull 10 hard positive samples from the VIS–NIR patch dataset, and then measure their similarity scores from the Siamese network, the naïve approach and SPIMNet. The results show that the Siamese network fails to predict the associations between all the samples. However, the naïve approach and SPIMNet exhibit much more accurate matching results than the Siamese network.

3.3. Loss function

To optimize SPIMNet, we use a loss function $L_{SPIMNet}$ as follows:

$$L_{SPIMNet} = L_m + y \times L_c, \tag{1}$$

where L_m helps SPIMNet to learn a determination of a similarity level between image patches. L_c encourages the domain conversion networks to translate images from one domain to another domain and vice versa.

For L_m, the similarity loss, we use the binary cross-entropy function as follows:

$$L_m = \hat{y}\log(y) + (1 - \hat{y})\log(1 - y), \tag{2}$$

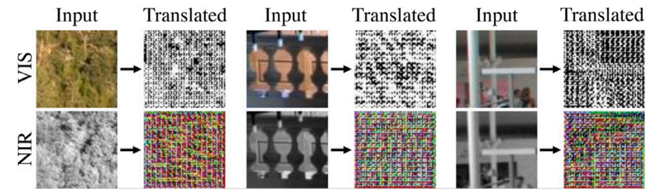


Fig. 6. Qualitative results of the VIS2NIR and NIR2VIS conversion networks without L_c term in a loss function.

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

Category	Number	Category	Number	Category	Number
Country	277,504	Field	240,896	Forest	376,832
Indoor	60,672	Mountain	151,296	Building	101,376
Street	164,608	Urban	147,712	Water	143,104

where $\hat{y}$ is the SPIMNet output for one training image and y is the class of the training example, i.e., $y = 1$ if it is a positive image pair and $y = 0$ if otherwise.

L_c, domain conversion loss, is calculated by outputs of each domain conversion network. The loss is a combination of perceptual loss [49] and L_1 loss. Both the perceptual loss and L_1 loss force the similarity between the input images for VIS2NIR and the output images from NIR2VIS, and vice versa. For the perceptual loss, we use a pre-trained VGG19 network ϕ [50] on the ImageNet dataset [51]. Because this pre-trained network generates well-distinguished feature, it is suitable for the perceptual loss. Let x_{vis} and x_{nir} be two input patches from two visible and NIR domains, respectively. The L_c is then computed as follows:

$$L_c = \alpha\left(\left|x_{vis} - x_{tvis}\right| + \left|x_{nir} - x_{tnir}\right|\right) + \beta\left(\left\|\phi(x_{vis}) - \phi(x_{tvis})\right\|_2^2 + \left\|\phi(x_{nir}) - \phi(x_{tnir})\right\|_2^2\right), \tag{3}$$

where x_{tvis} = NIR2VIS(x_{nir}) and x_{tnir} = VIS2NIR(x_{vis}). We set $\alpha = 0.1$ and $\beta = 30$ for all our experiments.

While L_m is a compulsory component of the patch matching task, we need to evaluate the effectiveness of the L_c. We train SPIMNet without L_c using the VIS–NIR patch dataset [16]. As shown in Fig. 6, the VIS2NIR and NIR2VIS conversion networks do not converge without L_c. The appearance of the translated images in Fig. 6 do not have enough informative features to distinguish an object or part of an object.

3.4. Training

We train our model from scratch for 35 epochs in total. All of the convolution and convolution transpose layers use the initialization method in [52] to set initial values for their weights. All models are trained in an end-to-end manner with the ADAM optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [53]. We use a batch size of 32 and set the learning rate to 0.0001 with a decay factor of 0.1 after 20 epochs. The training is performed with a customized version of Tensorflow 2.0 on an NVIDIA Titan Xp GPU, which usually takes two days. A forward pass of the proposed network takes about 13 ms for matched patches with 64 × 64 resolution.

To prevent an overfitting problem, all samples are normalized to [−1, 1], and data augmentation is carried out through random flipping, random rotation (90, 180, 270 degrees), and random cropping. In addition, two regularization techniques are employed: label smoothing [54] and L_2 kernel regularization for the convolution layers of the feature extraction networks with $l_2 = 0.001$.

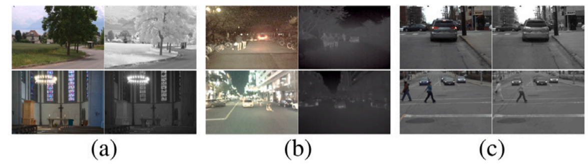


Fig. 7. Sample images of datasets used. (a) VIS–NIR patch dataset. (b) KAIST Multi-spectral pedestrian dataset. (c) PittsStereo-RGBNIR dataset.

4. Experimental results

To demonstrate the effectiveness of SPIMNet, we evaluate it on three publicly available datasets, the VIS–NIR patch dataset [16], the KAIST Multi-spectral pedestrian dataset [9], and the PittsStereo-RGBNIR dataset [45] as shown in Fig. 7. We compare SPIMNet with four hand-crafted feature matching methods (SIFT [12], GISIFT [55], EHD [56], LGHD [57]) and eight CNN-based methods including a Siamese network [16], Pseudo-Siamese network [16] (PSiamese), 2-channel network [16], PNNet [58], Q-Net [59], L2-Net [60], Hard-Net [61], SCFDM [15], and AFD-Net [14]. The false-positive rate in 95% recall (FPR95) is employed as a metric to evaluate the matching performance [14–16,60]. A smaller FPR95 represents better performance. In the KAIST Multi-spectral pedestrian dataset and the PittsStereo-RGBNIR dataset, we also measure FPR97/99 that is harder than FPR95.

4.1. VIS-NIR patch dataset

A public VIS–NIR scene dataset is introduced in [16] which contains visible and corresponding NIR images. The VIS–NIR patch dataset is currently used as a benchmark for evaluating descriptor learning and metric learning methods. The VIS–NIR patch dataset contains over 1.6 million patch-pairs, divided into 9 categories. The resolution of each patch is 64 × 64 pixels. The dataset providers extracted positive patch pairs using corresponding SIFT points between the visible and NIR images, while the negative patch pair is formed using a randomly selected patch in a NIR image over a patch in a visible image. Similar to [14,16,59], in our experiments, the Country category is only used for the training phase, and the remaining are used for testing. Table 1 shows the name and the number of samples for each category.

We demonstrate the effectiveness of SPIMNet on cross-spectral image patch matching by evaluating it and state-of-the-art methods as shown in Table 2. All methods are trained on the country category and tested on the other eight categories.

In general, the CNN-based methods perform better than the hand-crafted feature methods. The Siamese, PSiamese and 2-channel networks work well, even though they are not built for cross-spectral patch matching; SCFDM and AFD-Net, which are designed for cross-spectral image matching, show better performance than the hand-crafted feature methods. However, their direct extraction of discriminative features from input image patches means that the SCFDM and AFD-Net performances are limited.

SPIMNet divides the cross-spectral image matching into two independent tasks: domain conversion and similarity computation. This makes the problems tractable and permits the best performances in some categories. As shown in Fig. 4, the domain conversion part is the main difference between the proposed method and others. Other methods extract features from different domains and directly compare them in the metric learning module. However, our method converts one domain input to another domain. Therefore, the metric learning module can get features from the same domain as inputs, and comparing tasks becomes much easier. The effect of domain conversion can be found in Section 4.4.

In addition, we test two more setups for practical applications. First, we train our method for other patch sizes, 32 × 32 and 96 × 96. For

Table 2
A comparison of FPR95 among SPIMNet and 13 state-of-the-art methods on VIS–NIR patch dataset. All deep learning-based methods utilize data augmentation techniques. Lower is better, the best is bold and the second best is underlined.

Methods	Models	Field	Forest	Indoor	Mountain	Building	Street	Urban	Water
Traditional methods	SIFT [12]	39.44	11.39	10.13	28.63	19.69	31.14	10.85	40.33
	GISIFT [55]	34.75	16.63	10.63	19.52	12.54	21.80	7.21	25.75
	EHD [56]	33.85	19.61	24.23	26.32	17.11	22.31	3.77	19.80
	LGHD [57]	16.52	3.78	7.91	10.66	7.91	6.55	7.21	12.76
Descriptor learning	PN-Net [58]	20.09	3.27	6.36	11.53	5.19	5.62	3.31	10.72
	Q-Net [59]	17.01	2.70	6.16	9.61	4.61	3.99	2.83	8.44
	L2-Net [60]	16.77	0.76	2.07	5.98	1.89	2.83	0.62	11.11
	HardNet [61]	10.89	0.22	1.87	3.09	1.32	1.30	1.19	2.54
Metric learning	Siamese [16]	15.79	10.76	11.60	11.15	5.27	7.51	4.60	10.21
	PSiamese [16]	17.01	9.82	11.17	11.86	6.75	8.25	5.65	12.04
	2-channel [16]	9.96	0.12	4.40	8.89	2.30	2.18	1.58	6.40
	SCFDM [15]	7.91	0.87	3.93	5.07	2.27	2.22	0.85	4.75
	AFD-Net [14]	3.47	0.08	1.48	0.68	0.71	0.42	0.29	1.48
	MFD-Net [38]	2.59	**0.02**	**1.24**	0.95	0.48	0.24	**0.12**	**1.44**
Ours (SPIMNet)	32 × 32 patch	4.19	0.62	7.40	2.88	2.75	1.97	1.71	2.67
	64 × 64 patch	2.28	0.09	1.62	0.88	0.69	0.29	0.42	2.26
	96 × 96 patch	**1.76**	0.04	1.41	**0.31**	**0.44**	**0.14**	0.30	2.38
	Poor-aligned	2.49	0.13	3.85	0.93	0.91	0.41	0.66	2.74

all cases of the patch size, SPIMNet achieves outstanding performance and shows better performance when we use larger patch size images. Second, we make a poor-aligned dataset. The VIS–NIR dataset provides well-aligned data, but it is hard to get well-aligned data in practical applications. Some poor-aligned data will be mixed in the dataset if we gather data in practice. To show the robustness of our method, we add distortion(rotation:1° ∼10°, overlap: 90%∼100%) to half of the dataset and generate a poor-aligned dataset. Our method also shows less performance decrease in the poor-aligned dataset. These results can be found in Table 2.

4.2. KAIST Multi-spectral pedestrian dataset

The KAIST Multi-spectral pedestrian dataset [9] contains 95k visible–thermal image pairs of 12 sequences for road-driving scenes, and each image has a resolution of 640 × 480 pixels. The visible and thermal image pairs are physically aligned using beam splitter-based hardware. In this experiment, we split the first six sequences as training sets, and the remaining six sequences are used as test sets.

Similar to the dataset generation in Section 4.1, we crop the image patch to 64 × 64 and use a batch size of 8. For negative sample pairs, we randomly select a point in a thermal image for a point in a color image. In this way, we obtain over 1.1 million positive and negative sample pairs. Table 3 shows the number of samples in the training and test splits.

Table 3
The number of patch pairs in the training and test splits. The pairs were extracted from the KAIST multi-spectral pedestrian dataset.

Training	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
1,192,224	307,296	192,192	191,136	82,368	210,144	84,480

Since previous CNN-based methods in [16] do not cover visible–thermal matching and do not release their source codes, we implement and train the Siamese, PSiamese, and 2-channel networks on the dataset. The hyper-parameters of these three methods are used as described in the original paper.

Table 4 shows the quantitative results of the Siamese, PSiamese, 2-channel networks and SPIMNet. Compared to other methods, SPIMNet shows promising results on all sequences. We observe that the domain conversion network in SPIMNet translates images well from the visible domain to the thermal domain and vice versa, which leads to better discriminative feature learning than the direct feature extraction in [16].

4.3. PittsStereo-RGBNIR dataset

The PittsStereo-RGBNIR dataset [45] captures a 13.7-hour video using a vehicle-mounted VIS–NIR stereo system around the city of Pittsburgh. Each image has a 582 × 429 resolution. This dataset does not provide ground-truth correspondences. Accompanying this dataset, an unsupervised stereo matching is provided. We use the stereo matching to compute disparity maps on the dataset and filter out unreliable estimations via a left–right consistency check. We extract correspondences based on the disparity maps for valid pixels. In total, we obtain 109,146 sample patches for training and 4406 sample patches for testing. We train our model in this dataset for 60 epochs. We use a batch size of 32 and set the learning rate to 0.0001 with a decay factor of 0.1 after 30 epochs.

We train the Siamese, PSiamese, 2-channel, and SPIMNet using the training split from scratch. Table 5 shows quantitative results from the test split. In the FPR95 measurement, Siamese, PSiamese, and 2-channel networks show reasonable performances. However, when the thresholds of the error metric are set to FPR97 and FPR99, the comparison methods show more performance drops than the proposed method. For all metrics, SPIMNet shows higher performance. As shown in Table 5, when we implement other methods to compare with our method, we use a similar number of parameters for a fair comparison.

4.4. Ablation studies and additional analyses

An extensive ablation study is conducted to examine the effects of different components on SPIMNet performance.

First, we investigate the effect of the domain conversion loss (w/o L_c) and domain conversion networks (w/o 1DC and 2DC) in SPIMNet using the VIS–NIR patch dataset. The quantitative results using the FPR95 metric are shown in Table 6, indicating their performances are worse than SPIMNet. When L_c is not applied, we demonstrate the effectiveness of the automatic learned domain conversion as illustrated in Fig. 6. We only use the L_m term as a loss function on the same architecture. The network automatically converts data for comparison with other domain data when it is trained without the L_c, and it looks like it is working well. The performance is better than AFD-Net; SPIMNet, without the loss, can also take advantage of the domain conversion network. However, as shown in Fig. 6, it is just trained to generate one of the various solutions. The loss, L_c, can lead the training procedure to an obvious and reasonable solution, as shown in Fig. 3. It helps the network to be trained better, and the performance slightly increases. In particular, SPIMNet without the domain conversion networks (−2DC) suffers from significant performance drops, even worse than AFD-Net

Table 4
Quantitative results on KAIST multi-spectral pedestrian dataset. We use common quantitative measures of image matching with different spectral domains: FPR95 and FPR99 (Compared methods from [16]).

	Models	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
FPR95	Siamese	1.36	0.64	1.25	0.06	0.11	0.21
	PSiamese	1.91	1.17	1.36	0.32	0.17	0.81
	2-channel	0.75	0.57	1.18	0.03	0.02	0.09
	SPIMNet	**7.1e−3**	**0.01**	**0.02**	**1.7e−3**	**1.5e−3**	**5.6e−3**
FPR99	Siamese DA	2.27	1.32	2.9	0.19	0.37	0.49
	PSiamese	3.28	4.92	4.57	0.65	0.57	1.77
	2-channel	1.53	1.29	2.76	0.13	0.12	0.31
	SPIMNet	**0.04**	**0.08**	**0.08**	**0.03**	**0.06**	**0.07**

Table 5
Quantitative results of Siamese, PSiamese, 2-channel and SPIMNet on a PittsStereo-RGBNIR dataset. The used error metrics are FPR95, FPR97, and FPR99. The size column means the number of parameters for each network.

Models	FPR95	FPR97	FPR99	Size
Siamese [16]	0.01	0.02	0.05	38M
PSiamese [16]	2.7e−3	7.3e−3	0.04	143M
2-channel [16]	2.7e−3	8.4e−3	0.04	139M
SPIMNet	**4.5e−4**	**2.5e−3**	**0.02**	131M

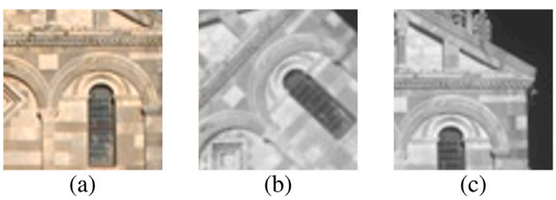


Fig. 8. Examples of geometric distortion. (a) Original image. (b) Rotated image. (c) Translated image.

and SCFDM. This validates the effectiveness of the domain conversion networks in SPIMNet.

Next, we check the performance of some geometric distortion data. We generate "overlap" data by translating patches a few pixels from the ground truth location and "rotation" data by rotating patches a few degrees. In overlapped data, the network should distinguish them from positive samples because we cropped patches from other locations. In rotated data, the network should generate similar results with positive samples because we cropped from the same locations. Because the similarity score from the last fully connected layer acts as distance in our metric learning framework, we compare this score to analyze the effect of geometric distortion. In Table 7, we report the result. Below 90% overlapped area(translated larger than 3 pixels in 64 × 64 patch), the proposed method generates distinguished results to positive samples. Also, in small rotations under 10 degrees, the network can match patches well. An example of the geometric distortion data is shown in Fig. 8.

Lastly, as another aspect of the analysis, we check the effective domain according to scene configurations. For this analysis, we compute the sum of the absolute differences in features between learned features from an input image and a translated image. The effective domain, which has a smaller difference, is chosen using the metric learning network. In Table 8, we report the ratio of the most effective domains for positive matching samples on the VIS–NIR patch dataset. We observe that the NIR–Pseudo NIR domain is more effective than the VIS–Pseudo VIS domain. However, the VIS–Pseudo VIS domain contributes to the discrimination as well.

4.5. Applications

To demonstrate the expandability of SPIMNet from some sensor fusion task, we conduct an additional experiment on cross-spectral stereo matching using 30 VIS–NIR image pairs. We build a cost volume for stereo matching with the output of SPIMNet. It looks like replacement original cost function to our SPIMNet, we select best scored points in search areas. Subsequently, we use the same post-processing techniques following [45] for a fair comparison with [25,29,45]. The quantitative results are reported in Table 9 whose examples are depicted in Fig. 9. This experiment shows that SPIMNet has the best performance over the comparison methods, proving its applicability for cross-spectral stereo matching.

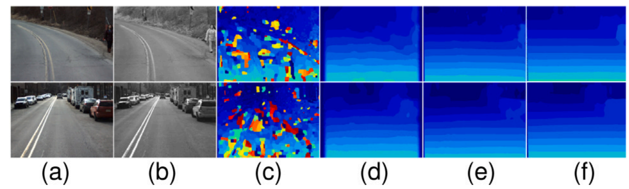


Fig. 9. Stereo matching results of SPIMNet and other methods. (a) VIS image. (b) NIR image. (c) ANCC [25]. (d) DASC [29]. (e) DMC [45]. (f) Ours.

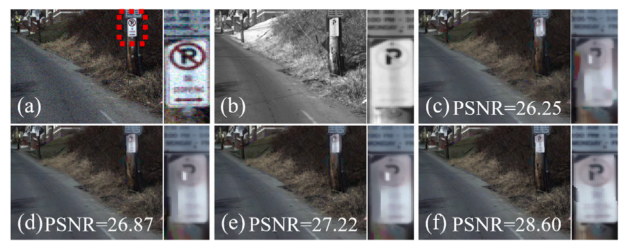


Fig. 10. An application to image enhancement. (a) Noisy VIS image. (b) Clean NIR image. (c) ANCC [25]. (d) DASC [29]. (e) DMC [45]. (f) Ours.

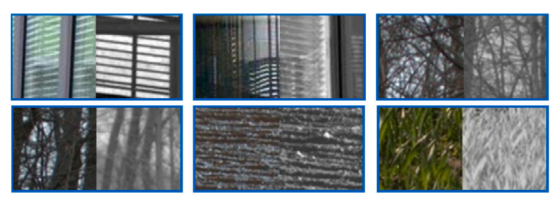


Fig. 11. Failure cases. The samples seem visually and semantically similar and SPIMNet classifies them as positive samples. However, the samples are labeled as negative.

In addition, SPIMNet is applicable for image enhancement. A noisy VIS and a clean NIR image are aligned by SPIMNet similarly to above mentioned cross-spectral stereo matching methods. Subsequently, we enhance the noisy image using a filtering process [62] whose guidance weights are based on intensity values of the NIR image. As shown in Fig. 10, we compare SPIMNet with other methods using PSNR metric for evaluating images enhancement performance, and SPIMNet shows better results.

Table 6
Ablation study on SPIMNet. Without the domain conversion loss (w/o L_c) and domain conversion networks (w/o 1DC: without NIR2VIS, w/o 2DC: without both networks). The best is bold and the second best is underlined.

Models	Field	Forest	Indoor	Mountain	Building	Street	Urban	Water
SPIMNet	**2.28**	**0.09**	**1.62**	0.88	**0.69**	0.29	**0.42**	**2.26**
w/o L_c	2.41	0.11	2.62	0.77	0.72	0.29	0.52	2.82
w/o 1DC	6.11	0.26	5.12	**0.66**	4.23	**0.28**	0.94	4.44
w/o 2DC	14.35	8.57	9.47	11.84	7.67	5.31	4.52	11.07

Table 7
Average scores of SPIMNet prediction with geometric distortion. (Original positive set: 0.95, Original negative set: 0.11)

Overlap(%)	100∼90	90∼80	80∼70	70∼60	60∼50
Score	0.94	0.80	0.48	0.24	0.15
Rotation(°)	0∼10	10∼20	20∼30	30∼40	40∼50
Score	0.93	0.70	0.28	0.01	0.05

Table 8
Ratio of the most effective domain for positive matching samples on the VIS–NIR patch dataset. (Unit: %)

Domain	Field	Forest	Indoor	Mountain
NIR-Pseudo NIR	63.8	63.9	60.2	61.3
VISL-Pseudo VIS	36.2	36.1	39.8	38.7
Domain	Building	Street	Urban	Water
NIR-Pseudo NIR	57.7	63.1	52.3	66.9
VISL-Pseudo VIS	42.3	36.9	47.7	33.1

Table 9
Quantitative comparison of stereo matching results using Root mean square error.

ANCC [25]	DASC [29]	DMC [45]	Ours
7.65	0.91	0.75	**0.54**

5. Conclusion

We have developed an image patch matching network across cross- and multi-spectral domains, named *SPIMNet*. SPIMNet is formulated as an end-to-end network, using two domain conversion networks to adjust the pixel- and intensity-level of input cross-spectral images. A dual-Siamese network enables the automatic selection of a better matching domain for two converted domain features. By incorporating these schemes in a deep learning framework, competitive matching accuracy is achieved on a variety of datasets, including visible–NIR and visible–thermal imagery. Many additional experiments have been conducted to show each module's robustness and effect. We have also shown that SPIMNet can be applied to various applications such as stereo matching and image enhancement.

Opportunities exist to improve SPIMNet as shown in Fig. 11 which reveal some of its failure cases. We observed that the number of false negatives is larger than false positive in inference of SPIMnet. It means that the negative samples are falsely classified as positive samples. Since the samples share the same semantic information and have similar appearances, this important challenge can be solved with an additional estimation of geometric features, such as 3D and surface normal information in an end-to-end learning framework.

CRediT authorship contribution statement

Yeongmin Ko: Methodology, Software, Validation, Writing – review & editing, Visualization. **Yong-Jun Jang:** Methodology, Software, Writing – review & editing. **Vinh Quang Dinh:** Conceptualization, Software, Writing – original draft. **Hae-Gon Jeon:** Formal analysis, Investigation, Writing – original draft. **Moongu Jeon:** Formal analysis, Investigation, Resources, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared our code link in the paper.

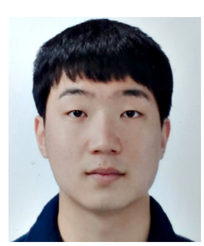
Acknowledgments

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. 2014-3-00077, Development of Global Multi-target Tracking and Event Prediction Techniques Based on Real-time Large-Scale Video Analysis).

References

[1] X. Jiang, J. Ma, G. Xiao, Z. Shao, X. Guo, A review of multimodal image matching: Methods and applications, Inf. Fusion 73 (2021) 22–71.
[2] Y. Choi, N. Kim, S. Hwang, K. Park, J.S. Yoon, K. An, I.S. Kweon, KAIST multi-spectral day/night data set for autonomous and assisted driving, IEEE Trans. Intell. Transp. Syst. (T-ITS) 19 (3) (2018) 934–948.
[3] C.A. Corneanu, M.O. Simón, J.F. Cohn, S.E. Guerrero, Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1548–1568.
[4] F. Abdullakutty, E. Elyan, P. Johnston, A review of state-of-the-art in Face Presentation Attack Detection: From early development to advanced deep learning and multi-modal fusion methods, Inf. Fusion 75 (2021) 55–69.
[5] T. Zhi, B.R. Pires, M. Hebert, S.G. Narasimhan, Multispectral imaging for fine-grained recognition of powders on complex backgrounds, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.
[6] P. Saponaro, S. Sorensen, A. Kolagunda, C. Kambhamettu, Material classification with thermal imagery, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[7] C. Wang, G. Yang, G. Papanastasiou, S.A. Tsaftaris, D.E. Newby, C. Gray, G. Macnaught, T.J. MacGillivray, DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis, Inf. Fusion 67 (2021) 147–160.
[8] C.S. Yilmaz, V. Yilmaz, O. Gungor, A theoretical and practical survey of image fusion methods for multispectral pansharpening, Inf. Fusion 79 (2022) 1–43.
[9] S. Hwang, J. Park, N. Kim, Y. Choi, I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[10] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, Z. Liu, Weakly aligned cross-modal learning for multispectral pedestrian detection, in: Proceedings of IEEE International Conference on Computer Vision, ICCV, 2019.
[11] D. Xu, W. Ouyang, E. Ricci, X. Wang, N. Sebe, Learning cross-modal deep representations for robust pedestrian detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[12] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[13] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (3) (2008) 346–359.
[14] D. Quan, X. Liang, S. Wang, S. Wei, Y. Li, N. Huyan, L. Jiao, AFD-Net: Aggregated feature difference learning for cross-spectral image patch matching, in: Proceedings of IEEE International Conference on Computer Vision, ICCV, 2019.

[15] D. Quan, S. Fang, X. Liang, S. Wang, L. Jiao, Cross-spectral image patch matching by learning features of the spatially connected patches in a shared space, in: Proceedings of Asian Conference on Computer Vision, ACCV, 2018.
[16] C.A. Aguilera, F.J. Aguilera, A.D. Sappa, C. Aguilera, R. Toledo, Learning cross-spectral similarity measures with deep convolutional neural networks, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[17] L. Zhou, Y. Ye, T. Tang, K. Nan, Y. Qin, Robust matching for SAR and optical images using multiscale convolutional gradient features, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.
[18] G.M. Dimitri, S. Spasov, A. Duggento, L. Passamonti, P. Lió, N. Toschi, Multi-modal and multicontrast image fusion via deep generative models, Inf. Fusion 88 (2022) 146–160.
[19] Y. Wu, J. Li, Y. Yuan, A.K. Qin, Q.-G. Miao, M.-G. Gong, Commonality autoencoder: Learning common features for change detection from heterogeneous images, IEEE Trans. Neural Netw. Learn. Syst. 33 (9) (2022) 4257–4270.
[20] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: Proceedings of European Conference on Computer Vision, ECCV, 2006.
[21] C. Liu, J. Yuen, A. Torralba, Sift flow: Dense correspondence across scenes and its applications, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 978–994.
[22] M. Brown, S. Susstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2011.
[23] T. Mouats, N. Aouf, A.D. Sappa, C. Aguilera, R. Toledo, Multispectral stereo odometry, IEEE Trans. Intell. Transp. Syst. 16 (3) (2015) 1210–1224.
[24] X. Shen, L. Xu, Q. Zhang, J. Jia, Multi-modal and multi-spectral registration for natural images, in: Proceedings of European Conference on Computer Vision, ECCV, 2014.
[25] Y.S. Heo, K.M. Lee, S.U. Lee, Robust stereo matching using adaptive normalized cross-correlation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (4) (2010) 807–822.
[26] Y.S. Heo, K.M. Lee, S.U. Lee, Joint depth map and color consistency estimation for stereo images with different illuminations and cameras, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2012) 1094–1106.
[27] P. Pinggera12, T. Breckon, H. Bischof, On cross-spectral stereo matching using dense gradient features, in: Proceedings of British Machine Vision Conference, BMVC, 2012.
[28] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2005.
[29] S. Kim, D. Min, B. Ham, M.N. Do, K. Sohn, DASC: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (9) (2016) 1712–1729.
[30] J. Holloway, K. Mitra, S.J. Koppal, A.N. Veeraraghavan, Generalized assorted camera arrays: Robust cross-channel registration and applications, IEEE Trans. Image Process. 24 (3) (2014) 823–835.
[31] M. Kim, D.K. Han, H. Ko, Joint patch clustering-based dictionary learning for multimodal image fusion, Inf. Fusion 27 (2016) 198–214.
[32] Y. Wu, J.-W. Liu, C.-Z. Zhu, Z.-F. Bai, Q.-G. Miao, W.-P. Ma, M.-G. Gong, Computational intelligence in remote sensing image registration: A survey, Int. J. Autom. Comput. 18 (1) (2021) 1–17.
[33] J. Zbontar, Y. LeCun, Stereo matching by training a convolutional neural network to compare image patches, J. Mach. Learn. Res. 17 (1) (2016) 2287–2318.
[34] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[35] P. Fischer, A. Dosovitskly, T. Brox, Descriptor matching with convolutional neural networks: a comparison to sift, 2014, arXiv preprint arXiv:1405.5769.
[36] X. Han, T. Leung, Y. Jia, R. Sukthankar, C.A. Berg, Matchnet: Unifying feature and metric learning for patch-based matching, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[37] D.-A. Beaupre, G.-A. Bilodeau, Domain siamese CNNs for sparse multispectral disparity estimation, in: 2020 25th International Conference on Pattern Recognition, ICPR, 2021, pp. 3667–3674.
[38] C. Yu, Y. Liu, C. Li, L. Qi, X. Xia, T. Liu, Z. Hu, Multibranch feature difference learning network for cross-spectral image patch matching, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.
[39] P. Yan, Y. Tan, Y. Tai, D. Wu, H. Luo, X. Hao, Unsupervised learning framework for interest point detection and description via properties optimization, Pattern Recognit. 112 (2021) 107808.
[40] Y. Ye, T. Tang, B. Zhu, C. Yang, B. Li, S. Hao, A multiscale framework with unsupervised learning for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15.
[41] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[42] K. Vougioukas, S. Petridis, M. Pantic, {DINO}: A conditional energy-based {GAN} for domain translation, in: International Conference on Learning Representations, 2021.
[43] Y. Wu, Z. Bai, Q. Miao, W. Ma, Y. Yang, M. Gong, A classified adversarial network for multi-spectral remote sensing image change detection, Remote Sens. 12 (13) (2020) 2098.
[44] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[45] T. Zhi, B.R. Pires, M. Hebert, S.G. Narasimhan, Deep material-aware cross-spectral stereo matching, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[46] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[47] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of International Conference on Machine Learning, ICML, 2015.
[48] D. Ulyanov, A. Vedaldi, V.S. Lempitsky, Instance normalization: The missing ingredient for fast stylization, 2016, CoRR.
[49] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of European Conference on Computer Vision, ECCV, 2016.
[50] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, ICLR, 2015.
[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2009.
[52] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification, in: Proceedings of IEEE International Conference on Computer Vision, ICCV, 2015.
[53] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, ICLR, 2014.
[54] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016.
[55] D. Firmenichy, M. Brown, S. Süsstrunk, Multispectral interest points for RGB-nir image registration, in: Proceedings of International Conference on Image Processing, ICIP, 2011.
[56] C. Aguilera, F. Barrera, F. Lumbreras, A.D. Sappa, R. Toledo, Multispectral image feature points, Sensors 12 (9) (2012) 12661–12672.
[57] C.A. Aguilera, A.D. Sappa, R. Toledo, LGHD: A feature descriptor for matching across non-linear intensity variations, in: Proceedings of International Conference on Image Processing, ICIP, 2015.
[58] V. Balntas, E. Johns, L. Tang, K. Mikolajczyk, PN-net: Conjoined triple deep network for learning local image descriptors, 2016, arXiv preprint arXiv:1601.05030.
[59] C.A. Aguilera, A.D. Sappa, C. Aguilera, R. Toledo, Cross-spectral local descriptors via quadruplet network, Sensors 17 (4) (2017).
[60] Y. Tian, B. Fan, F. Wu, L2-Net: Deep learning of discriminative patch descriptor in euclidean space, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017.
[61] A. Mishchuk, D. Mishkin, F. Radenovic, J. Matas, Working hard to know your neighbor's margins: Local descriptor learning loss, in: Advances in Neural Information Processing Systems (NeurIPS), 2017.
[62] X. Shen, C. Zhou, L. Xu, J. Jia, Mutual-structure for joint filtering, in: Proceedings of IEEE International Conference on Computer Vision, ICCV, 2015.



Yeongmin Ko received the B.S. degree in School of Electrical Engineering from Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology. His current research interests include computer vision, self-driving, and deep learning.

Yong-Jun Chang received the B.S. degree in electronic engineering and avionics from Korea Aerospace University, Gyeonggi-do, Korea, in 2014, and he received the M.S. degree in information and communications and the Ph.D degree in electrical engineering and computer science from Gwangju Institute of Science and Technology (GIST), Gwangju, Korea, in 2016 and 2021. In 2021, he was a researcher of Korea Culture Technology Institute in GIST. He is currently a research engineer of Hyundai Rotem. His research interests are in computer vision and deep learning.

Vinh Quang Dinh received the B.S. degree in computer science from Nong Lam University, Ho Chi Minh City, Vietnam, in 2008, and the M.S. and Ph.D. degrees in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2013 and 2016, respectively. From 2016 to 2017, he was a Postgraduate Researcher with Sungkyunkwan University. From 2017 to 2020, he was a Postgraduate Researcher with the Gwangju Institute of Science and Technology. In 2020, he joined Vietnamese-German University, where he is currently a Lecturer with the School of Electrical Engineering and Computer Science. His current research interests include computer vision and deep learning.

Hae-Gon Jeon received the BS degree in the School of Electrical and Electronic Engineering from Yonsei University in 2011, the MS degree and Ph.D. degree in the School of Electrical Engineering from KAIST in 2013 and in 2018, respectively. He was a postdoctoral researcher of the Robotics Institute at Carnegie Mellon University. He is currently affiliated with both AI Graduate School and the School of Electrical Engineering and Computer Science at GIST as an assistant professor. He is a winner of the Best Ph.D. Thesis Award 2018 in KAIST. His research interests include computational imaging, 3D reconstruction and machine learning.

Moongu Jeon received the B.S. degree in architectural engineering from Korea University, Seoul, South Korea, in 1988, and the M.S. and Ph.D. degrees in computer science and scientific computation from the University of Minnesota, Minneapolis, MN, USA, in 1999 and 2001, respectively. As the master's degree researcher, he was involved in optimal control problems with the University of California at Santa Barbara, Santa Barbara, CA, USA, from 2001 to 2003, and then moved to the National Research Council of Canada, where he was involved in the sparse representation of high-dimensional data and the image processing, until July 2005. In 2005, he joined the Gwangju Institute of Science and Technology, Gwangju, South Korea, where he is currently a Full Professor with the School of Electrical Engineering and Computer Science. His current research interests include machine learning, computer vision, and artificial intelligence.