Spectral-invariant Matching Network

Yeongmin Ko[a], Yong-Jun Jang[b], Vinh Quang Dinh[c], Hae-Gon Jeon[d] and 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, Vietnam,
[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:
RGB-NIR
image matching
image patch
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 state-of-the-art results on a variety of challenging datasets, including both RGB-NIR and RGB-Thermal image pairs. To demonstrate its robustness, we submit the source codes of SPIMNet as supplementary material.

1. Introduction

In the field of computer vision, cross-spectral (*i.e.*visible-near infrared (NIR)) and multi-spectral (*i.e.*visible-thermal) image matching is being actively studied because the different spectral domains can provide complementary information. 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 [1]. All-day vision has become an essential and significant task for sensor fusion systems that conduct facial expression recognition [2], material classification [3, 4], and pedestrian detection [5, 6, 7]. However, matching images across different domains is still a challenging problem.

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 [8, 9], 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 [10, 11, 12]. The most promising of these methods predicts similarities between two patches directly encoded from the common Siamese structures; however, we have observed that the Siamese structures are not suited to dealing with both intensity- and pixel-level differences.

In this paper, we present a SPectral-Invariant Matching Network (SPIMNet), an end-to-end CNN framework for robust image patch matching across different spectral domains. In contrast to previous methods that extract features directly from input patches, SPIMNet first learns the spectral translations of input patches using a domain conversion network. We then utilize a dual-Siamese network for feature extraction from each translated piece of information to predict the matching label through a fully connected network.

Using this end-to-end network to conduct image patch matching across different spectral domains, we obtain state-of-the-art results over several standard datasets including both visible-NIR and visible-thermal images. Ablation studies indicate that each of these technical contributions leads to appreciable improvements in matching accuracy.

2. Related Works

Our work is closely related to a similarity computation. Below, we describe related works in the areas of image matching across different spectral domains based on hand-crafted feature descriptors and CNNs.

Hand-crafted Feature Descriptions hand-crafted features such as SIFT [8], SURF [9] and FAST [13] 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 [14]. However, these methods often fail in cross- and multi-spectral imagery

*Corresponding author.
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)
ORCID(s): 0000-0002-2775-7789 (M. Jeon)

(a) Visible-NIR image pairs

(b) Visible-Thermal image pairs

Figure 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.

because their different spectral characteristics result in nonlinear variations in intensity, and inconsistent textures.

Alternative methods, such as the multi-spectral SIFT [15] improved the performance of scene category recognition for a pair of RGB-NIR scenes by analyzing the statistical dependencies between them. In [16], 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 [17], 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*. [18] analyzed a color intensity model and proposed an adaptive normalized cross-correlation for stereo matching, and their extended work [19] iteratively estimated dense depth maps and adjusted color consistency. In [20], cross-spectral stereo matching was presented with dense gradient features based on the HOG descriptor [21]. Kim *et al*. [22] proposed a dense descriptor for cross-spectral correspondences with their adaptive self-correlation and randomized receptive field pooling. Holloway *et al*. [23] presented an assorted camera array and a normalized gradient cost to measure correspondences in cross-spectral images.

CNN-based Multi-spectral Image Matching CNNs have achieved great success in image patch matching, thereby significantly improving state-of-the-art stereo matching [24] and similarity computation [25]. In particular, Siamese structures have demonstrated robustness when performing image matching for various datasets in [26]. A generalization of the Siamese structure in [27] showed promising performance over hand-crafted features by simultaneously learning local patch representation and performing feature comparisons.

Aguilera *et al*. [12] trained a Siamese network for multi-spectral image matching. Quan *et al*. [11] 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 [10].

Nalla *et al*. [28] applied the cross-spectral matching method using a domain adaptation of iris recognition. In [29, 30], multi-spectral images are utilized for person re-identification. Lu *et al*. [29] proposed two stream networks for feature extraction from multi-spectral images, and fused them with a shared-specific feature transfer. Ye *et al*. [30] enhanced a performance of cross-modal person re-identification considering both an intra-modality weighted-part aggregation and a cross-modality graph structured attention. Zhi *et al*. [31] presented a weakly supervised learning framework for dense depth computation from visible and NIR image pairs. We note that the work in [31] 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 [31], the domain conversion in SPIMNet learns to make a translated feature prior to encoding the image feature. In Sec.4, we will demonstrate the effectiveness of this domain conversion for multi-spectral patch matching.

3. Spectral-invariant Matching Network

Previous works [10, 11] 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 appearances. For this reason, performance has been limited in previous works [10, 11].

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 Sec.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 [32] with ten blocks. The encoding blocks consist of convolution, batch norm [33], instance norm [34], 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

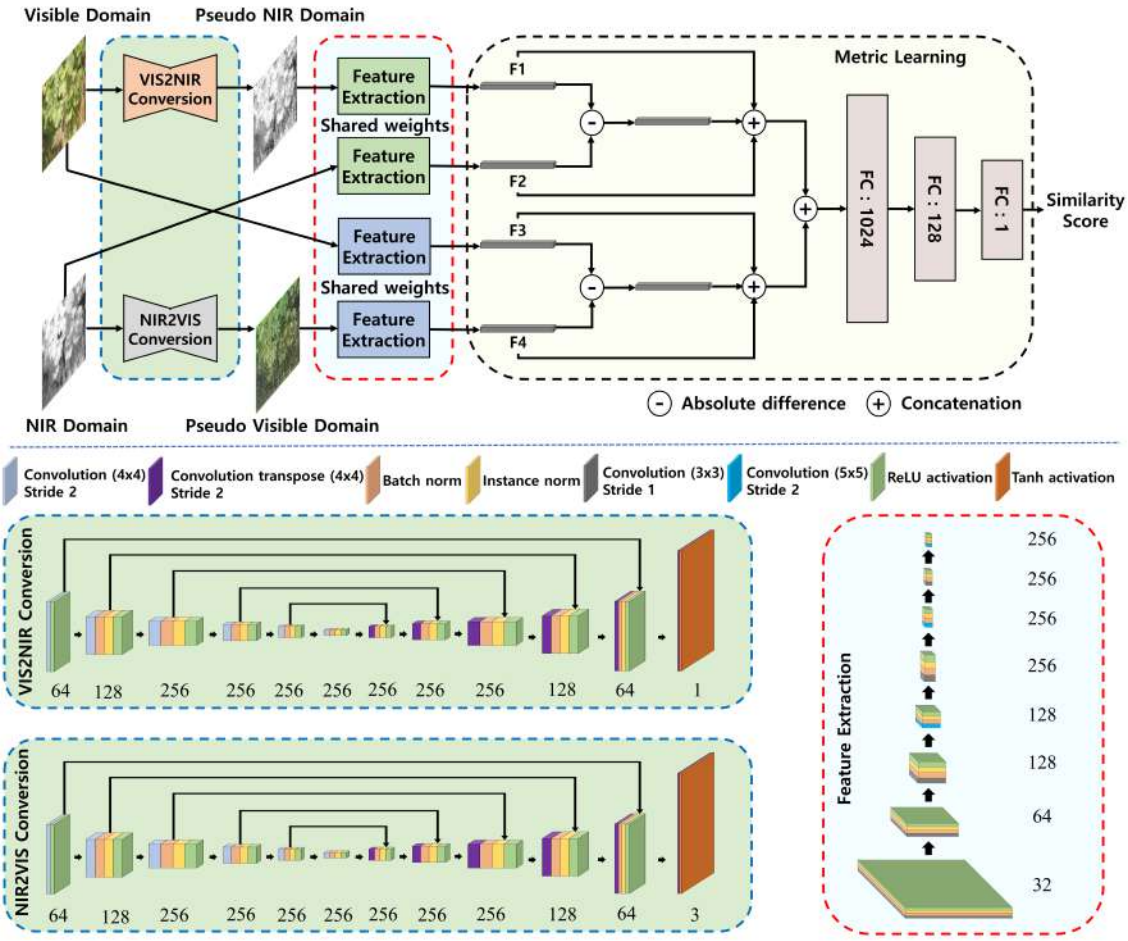


Figure 2: An overview of SPIMNet and details of its sub-networks. SPIMNet consists of two domain conversion networks (VIS2NIR and NIR2VIS for cross-spectral image matching), two feature extraction networks and a metric learning network. The feature extraction networks extract discriminative features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

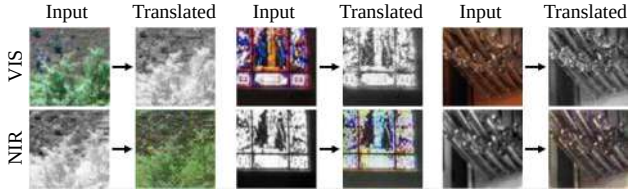


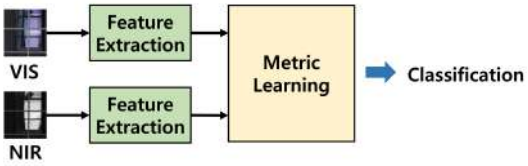
Figure 3: Qualitative results of the VIS2NIR and NIR2VIS conversion network using the VIS-NIR patch dataset. The left column means input images, and the right columns are translated images.

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].

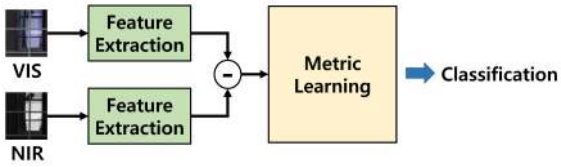
Fig.3 shows examples of VIS2NIR and NIR2VIS using the VIS-NIR patch dataset [12]; details are described in Sec.4.1. The translated images not only keep the low-level features, but also are similar in appearance to their corresponding input images. Although the intensity levels between the input visible 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.

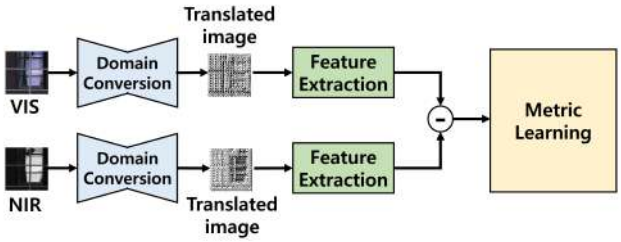
The feature extraction network is comprised of eight layers, each of which includes a convolution, batch norm, instance norm, and ReLU activation. The number of filters, filter size, and stride for the convolution layers of the eight blocks are (32, 3×3, 1), (64, 3×3, 1), (128, 3×3, 1), (128, 5×5, 2), (256, 3×3, 1), (256, 5×5, 2), (256, 3×3, 1) and (256, 5×5, 2), respectively.



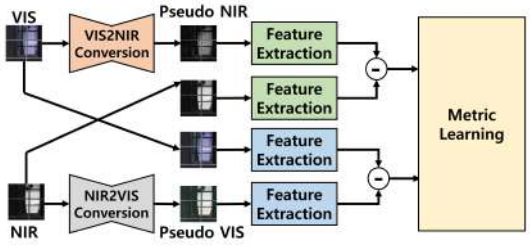
(a) Siamese Network [12]



(b) AFD-Net [10]



(c) Naïve approach



(d) SPIMNet

Figure 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.

Metric Learning Network Finally, the extracted features are fed into a metric learning network which includes fully connected layers. 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 [12]. The network learns to determine the similarities between two patches. AFD-Net [10] 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.

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,

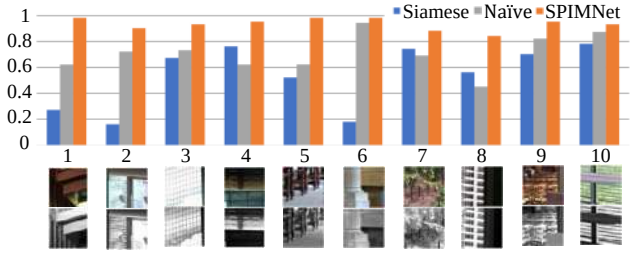


Figure 5: Similarity scores of Siamese network, naïve approach, and SPIMNet for the 10 hard positive samples.

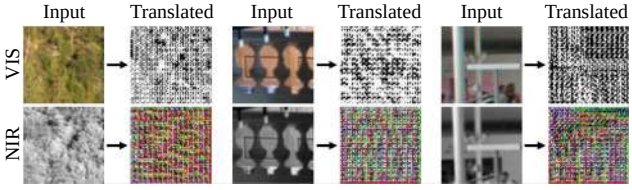


Figure 6: Qualitative results of the VIS2NIR and NIR2VIS conversion networks without L_c term in a loss function.

SPIMNet can significantly 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, we use the binary cross-entropy function as follows:

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

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 is a combination of perceptual loss [35] 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 pretrained VGG19 network ϕ [36] on the ImageNet dataset [37]. 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:

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

where $x_{tvis} = \text{NIR2VIS}(x_{nir})$ and $x_{tnir} = \text{VIS2NIR}(x_{vis})$. 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 RGB-NIR patch dataset [12]. 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.

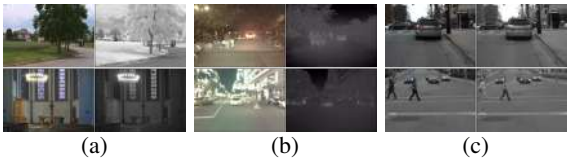
3.4. Training

We train our model from scratch for 35 epochs in total. All of the convolution and convolution transpose layers use the initialization method in [38] to set initial values for their weights. All models are trained in an end-to-end manner with the ADAM optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) [39]. 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 milliseconds 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 [40] and L_2 kernel regularization for the convolution layers of the feature extraction networks with $l_2 = 0.001$.

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



(a) (b) (c)

Figure 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 [12], the KAIST Multi-spectral pedestrian dataset [5], and the PittsStereo-RGBNIR dataset [31] as shown in Fig.7. We compare SPIMNet with four handcrafted feature matching methods (SIFT [8], GISIFT [41], EHD [42], LGHD [43]) and eight CNN-based methods including a Siamese network [12], Pseudo-Siamese network [12] (PSiamese), 2-channel network [12], PNNet [44], Q-Net [45], L2-Net [46], HardNet [47], SCFDM [11], and AFD-Net [10]. The false-positive rate in 95% recall (FPR95) is employed as a metric to evaluate the matching performance [12, 11, 10, 46]. A smaller FPR95 represents better performance.

4.1. VIS-NIR patch dataset

A public VIS-NIR scene dataset is introduced in [12] 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. A positive patch pair is extracted 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 [12, 45, 10], 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

Table 2
A comparison of FPR95 among SPIMNet and 13 state-of-the-art methods on VIS-NIR patch dataset. DA denotes data augmentation in the training process. Except for SPIMNet, the evaluation results are directly sourced from [10] for a fair comparison. The best performance is written in bold.

Methods	Models	Field	Forest	Indoor	Mountain	Building	Street	Urban	Water
Traditional methods	SIFT [8]	39.44	11.39	10.13	28.63	19.69	31.14	10.85	40.33
	GISIFT [41]	34.75	16.63	10.63	19.52	12.54	21.80	7.21	25.75
	EHD [42]	33.85	19.61	24.23	26.32	17.11	22.31	3.77	19.80
	LGHD [43]	16.52	3.78	7.91	10.66	7.91	6.55	7.21	12.76
Descriptor learning	PN-Net DA [44]	20.09	3.27	6.36	11.53	5.19	5.62	3.31	10.72
	Q-Net DA [45]	17.01	2.70	6.16	9.61	4.61	3.99	2.83	8.44
	L2-Net DA [46]	16.77	0.76	2.07	5.98	1.89	2.83	0.62	11.11
	HardNet DA [47]	10.89	0.22	1.87	3.09	1.32	1.30	1.19	2.54
Metric learning	Siamese DA [12]	15.79	10.76	11.60	11.15	5.27	7.51	4.60	10.21
	PSiamese DA [12]	17.01	9.82	11.17	11.86	6.75	8.25	5.65	12.04
	2-channel DA [12]	9.96	0.12	4.40	8.89	2.30	2.18	1.58	6.40
	SCFDM DA [11]	7.91	0.87	3.93	5.07	2.27	2.22	0.85	4.75
	AFD-Net DA [10]	3.47	0.08	1.48	0.68	0.71	0.42	0.29	1.48
	SPIMNet DA	**0.02**	**7.0e-4**	**0.02**	**7.4e-3**	**5.3e-3**	**2.4e-3**	**3.3e-3**	**0.02**

trained on the country category and tested on the other eight categories. SPIMNet significantly outperforms the other methods in all test 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 all categories.

4.2. KAIST Multi-spectral pedestrian dataset

The KAIST Multi-spectral pedestrian dataset [5] 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 Sec.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.

Since previous CNN-based methods in [12] 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 [12].

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

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 [12]).

	Models	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
FPR95	Siamese DA	1.36	0.64	1.25	0.06	0.11	0.21
	PSiamese DA	1.91	1.17	1.36	0.32	0.17	0.81
	2-channel DA	0.75	0.57	1.18	0.03	0.02	0.09
	SPIMNet DA	**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 DA	3.28	4.92	4.57	0.65	0.57	1.77
	2-channel DA	1.53	1.29	2.76	0.13	0.12	0.31
	SPIMNet DA	**0.04**	**0.08**	**0.08**	**0.03**	**0.06**	**0.07**

4.3. PittsStereo-RGBNIR Dataset

The PittsStereo-RGBNIR dataset [31] 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

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.

Models	FPR95	FPR97	FPR99
Siamese DA [12]	0.01	0.02	0.05
PSiamese DA [12]	2.7e-3	7.3e-3	0.04
2-channel DA [12]	2.7e-3	8.4e-3	0.04
SPIMNet DA	**4.5e-4**	**2.5e-3**	**0.02**

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 4,406 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.

4.4. Ablation Studies

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

First, we investigate the effect of data augmentation (DA), batch norm (BN), and instance norm (IN), regularization techniques including label smoothing and L_2 kernel regularization (RE), and domain conversion networks 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. This validates the effectiveness of each component in SPIMNet. In particular, SPIMNet without the domain conversion networks (-2DC) suffers from significant performance drops, even worse than ADF-Net and SCFDM.

In addition, we compare SPIMNet with a naïve approach that has one feature extraction and two domain conversion networks as illustrated in Fig.4(c). In this experiment, we demonstrate the effectiveness of the automatic learned feature selection. In the naïve approach, the two domain conversion networks have the same architecture as the VIS2NIR network. We only use the L_m term as a loss function to force the converted images into one common domain. As shown in Table 7, the performance of SPIMNet is higher than that of the naïve approach, as expected, because embedding both the F1-F2 and F3-F4 features allow SPIMNet to automatically select a matching domain. Another interesting point is that the naïve approach shows better performance than AFD-Net. We observe that the naïve approach can also take advantage of the domain conversion network.

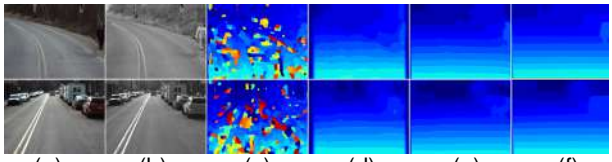


Figure 8: Stereo matching results of SPIMNet and other methods. (a) RGB image. (b) NIR image. (c) ANCC [18]. (d) DASC [22]. (e) DMC [31]. (f) Ours

Figure 9: An application to image enhancement. (a) Noisy RGB. (b) Clean NIR. (c) ANCC [18]. (d) DASC [22]. (e) DMC [31]. (f) Ours.

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, we conduct an additional experiment on cross-spectral stereo matching using 30 RGB-NIR image pairs. We build a cost volume for stereo matching with the output of SPIMNet. Subsequently, we use the same post-processing techniques following [31] for a fair comparison with [18, 31, 22]. The quantitative results are reported in Table 9 whose examples are depicted in Fig.8. This experiment shows that SPIMNet has the best performance over the comparison methods, proving its applicability for cross-spectral stereo matching.

In addition, SPIMNet is applicable for image enhancement. A noisy RGB and a clean NIR image is aligned by searching for correspondences with SPIMNet and above mentioned cross-spectral stereo matching methods. Subsequently, we enhance the noisy image using a filtering process [48] whose guidance weights are based on intensity values of the NIR image. As shown in Fig.9, SPIMNet produces accurate correspondences even with severe noise in the RGB image.

Table 6
Ablation study on SPIMNet. Without data augmentation (DA), batch norm (BN), instance norm (IN), L_2 kernel regularization (RE) and domain conversion networks (VIS2NIR and NIR2VIS). Bold: Best, Underbar: Second best.

Models	Field	Forest	Indoor	Mountain	Building	Street	Urban	Water
SPIMNet	**0.05**	**5.3e-4**	**0.43**	9.3e-3	**0.02**	**0.0**	**5.4e-3**	**0.03**
w/o DA	0.06	5.4e-4	0.64	**1.5e-4**	0.04	6.1e-7	8.9e-3	0.06
w/o BN	3.51	0.31	5.58	1.27	1.81	0.46	1.01	1.58
w/o IN	2.68	0.18	3.63	0.94	0.87	0.16	0.71	1.39
w/o RE	**0.05**	9.6e-4	0.49	0.01	0.03	3.6e-6	0.01	0.04
w/o 1DC	0.06	2.6e-3	0.51	6.6e-3	0.04	2.8e-3	9.4e-3	0.04
w/o 2DC	14.35	8.57	9.47	11.84	7.67	5.31	4.52	11.07

Table 7
Quantitative results of SPIMNet and the naïve approach on the VIS-NIR patch dataset.

Models	Field	Forest	Indoor	Mountain	Building	Street	Urban	Water
SPIMNet	0.05	5.3e-4	0.43	9.3e-3	0.02	0.0	5.4e-3	0.03
Naïve	0.1	0.01	0.62	0.01	0.07	3.1e-3	0.01	0.06

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

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

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

ANCC [18]	DASC [22]	DMC [31]	Ours
7.65	0.91	0.75	**0.54**

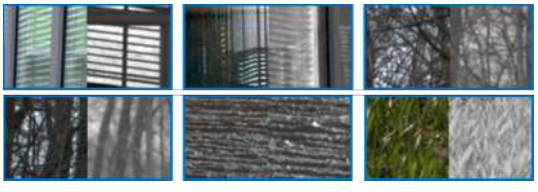


Figure 10: Failure cases. The samples seem visually and semantically similar and SPIMNet classifies them as positive samples. However, the samples are labelled as negative.

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, state-of-the-art matching accuracy is achieved on a variety of datasets, including visible-NIR and visible-thermal imagery.

Opportunities exist to improve SPIMNet as shown in Fig.10 which reveal some of its failure cases. Most of the wrong classifications are false negatives, meaning 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.

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] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, vol. 19, no. 3, pp. 934–948, 2018.

[2] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 38, no. 8, pp. 1548–1568, 2016.

[3] T. Zhi, B. R. Pires, M. Hebert, and 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.

[4] P. Saponaro, S. Sorensen, A. Kolagunda, and C. Kambhamettu, "Material classification with thermal imagery," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019.

[7] D. Xu, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Learning cross-modal deep representations for robust pedestrian detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding (CVIU)*, vol. 110, no. 3, pp. 346–359, 2008.

[10] D. Quan, X. Liang, S. Wang, S. Wei, Y. Li, N. Huyan, and 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.

[11] D. Quan, S. Fang, X. Liang, S. Wang, and 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.

[12] C. A. Aguilera, F. J. Aguilera, A. D. Sappa, C. Aguilera, and 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.

[13] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2006.

[14] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 5, pp. 978–994, 2011.

[15] M. Brown and S. Susstrunk, "Multi-spectral sift for scene category recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[16] T. Mouats, N. Aouf, A. D. Sappa, C. Aguilera, and R. Toledo, "Multispectral stereo odometry," *IEEE Transactions on Intelligent Transportation Systems (T-ITS)*, vol. 16, no. 3, pp. 1210–1224, 2015.

[17] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multispectral registration for natural images," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[18] Y. S. Heo, K. M. Lee, and S. U. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 4, pp. 807–822, 2010.

[19] Y. S. Heo, K. M. Lee, and S. U. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 5, pp. 1094–1106, 2012.

[20] P. Pinggera12, T. Breckon, and H. Bischof, "On cross-spectral stereo matching using dense gradient features," in *Proceedings of British Machine Vision Conference (BMVC)*, 2012.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[22] S. Kim, D. Min, B. Ham, M. N. Do, and K. Sohn, "Dasc: Robust dense descriptor for multi-modal and multi-spectral correspondence estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 39, no. 9, pp. 1712–1729, 2016.

[23] J. Holloway, K. Mitra, S. J. Koppal, and A. N. Veeraraghavan, "Generalized assorted camera arrays: Robust cross-channel registration and applications," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 3, pp. 823–835, 2014.

[24] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research (JMLR)*, vol. 17, no. 1, pp. 2287–2318, 2016.

[25] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] P. Fischer, A. Dosovitskly, and T. Brox, "Descriptor matching with convolutional neural networks: a comparison to sift," 2014. arXiv preprint arXiv:1405.5769.

[27] X. Han, T. Leung, Y. Jia, R. Sukthankar, and 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.

[28] P. R. Nalla and A. Kumar, "Toward more accurate iris recognition using cross-spectral matching," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 1, pp. 208–221, 2017.

[29] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, "Dynamic dual-attentive aggregation learning for visible-infrared person reidentification," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[31] T. Zhi, B. R. Pires, M. Hebert, and S. G. Narasimhan, "Deep material-aware cross-spectral stereo matching," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), 2015.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

[34] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, 2016.

[35] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015.

[39] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2014.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

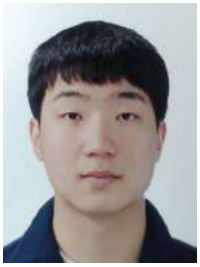
[41] D. Firmenichy, M. Brown, and S. Süsstrunk, "Multispectral interest points for rgb-nir image registration," in *Proceedings of International Conference on Image Processing (ICIP)*, 2011.

[42] C. Aguilera, F. Barrera, F. Lumbreras, A. D. Sappa, and R. Toledo, "Multispectral image feature points," *Sensors*, vol. 12, no. 9, pp. 12661–12672, 2012.

[43] C. A. Aguilera, A. D. Sappa, and R. Toledo, "Lghd: A feature descriptor for matching across non-linear intensity variations," in *Proceedings of International Conference on Image Processing (ICIP)*, 2015.

[44] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk, "Pn-net: Conjoined triple deep network for learning local image descriptors," *arXiv preprint arXiv:1601.05030*, 2016.

[45] C. A. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Cross-spectral local descriptors via quadruplet network," *Sensors*, vol. 17, no. 4, 2017.
[46] Y. Tian, B. Fan, and 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.
[47] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
[48] X. Shen, C. Zhou, L. Xu, and 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.