Information Fusion 94 (2023) 200–214

Contents lists available at ScienceDirect

Information Fusion

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

Full length article

PCNet: A structure similarity enhancement method for multispectral and multimodal image registration

Si-Yuan Cao [a,b], Beinan Yu [b], Lun Luo [b], Runmin Zhang [b], Shu-Jie Chen [d], Chunguang Li [b], Hui-Liang Shen [b,c,*]

[a] Ningbo Innovation Center, Zhejiang University, Ningbo, 315100, Zhejiang, China
[b] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, 310027, Zhejiang, China
[c] Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province, Hangzhou, 310015, Zhejiang, China
[d] School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, 310018, Zhejiang, China

ARTICLE INFO

Keywords:
Multispectral image
Multimodal image
Image registration
Phase congruency
Similarity enhancement
Convolutional neural network

ABSTRACT

Multispectral and multimodal images are of important usage in the field of multi-source visual information fusion. Due to the alternation or movement of image devices, the acquired multispectral and multimodal images are usually misaligned, and hence image registration is pre-requisite. Different from the registration of common images, the registration of multispectral or multimodal images is a challenging problem due to the nonlinear variation of intensity and gradient. To cope with this challenge, we propose the phase congruency network (PCNet) to enhance the structure similarity of multispectral or multimodal images. The images can then be aligned using the similarity-enhanced feature maps produced by the network. PCNet is constructed under the inspiration of the well-known phase congruency. The network embeds the phase congruency prior into two simple trainable layers and series of modified learnable Gabor kernels. Thanks to the prior knowledge, once trained, PCNet is applicable on a variety of multispectral and multimodal data such as flash/no-flash and RGB/NIR images without additional further tuning. The prior also makes the network lightweight. The trainable parameters of PCNet is 2400× and 1500×less than the deep-learning registration method deep homography network (DHN) and unsupervised deep homography network (UDHN), while its registration performance surpasses them. Experimental results validate that PCNet outperforms current state-of-the-art conventional multimodal registration algorithms. Besides, PCNet can act as a complementary part of the deep-learning registration methods, which significantly boosts their registration accuracy. On the Columbia imaging and vision laboratory (CAVE) multispectral dataset, the percentage of the number of images under 1 pixel average corner error (ACE) of UDHN is raised from 0.1% to 82.5% after the processing of PCNet.

1. Introduction

Multispectral and multimodal images, such as RGB and near-infrared (NIR) images [1], flash/no-flash images [2], and multispectral band images [3,4], usually contain much richer information compared to conventional RGB images. They are important data for multi-source visual information fusion applications including pedestrian detection and re-identification [5–7], image fusion [8–10], and image denoising/dehazing [11,12]. Pixel-level alignment of multispectral/multimodal images is a fundamental requirement for these tasks. Nevertheless, multispectral/multimodal images are prone to be misaligned due to the alternation or movement of image devices [3,13]. Therefore, registering multispectral/multimodal images is the primary problem for further computer vision and computational photography tasks.

The most intractable aspect of multispectral/multimodal image registration is the ubiquitous variation of intensity and gradient among data from different sources [13,14]. To cope with the challenge, various image registration methods have been proposed. The registration methods can be categorized into feature-based ones and intensity-based ones [15]. Feature-based methods align images through feature detection, feature description, and transformation estimation [16]. Intensity-based methods align images by finding the best correspondence maximizing (or minimizing) specific similarity measures between the two input images [17]. Various similarity measures have been specially developed for multispectral/multimodal images, however, they are usually of complicated forms, making their optimization problematic

* Corresponding author at: College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, 310027, Zhejiang, China.
E-mail addresses: karlcao@hotmail.com (S.-Y. Cao), mr_vernon@hotmail.com (B. Yu), luolun@zju.edu.cn (L. Luo), runmin_zhang@zju.edu.cn (R. Zhang), chenshujie@zjgsu.edu.cn (S.-J. Chen), cgli@zju.edu.cn (C. Li), shenhl@zju.edu.cn (H.-L. Shen).

https://doi.org/10.1016/j.inffus.2023.02.004
Received 10 October 2022; Received in revised form 6 January 2023; Accepted 1 February 2023
Available online 4 February 2023

and computationally expensive [18]. Another kind of intensity-based method aims to enhance the structural similarity between images by some transformations in advance and then register the transformed images using common measures such as cross correlation [19], sum of squared differences (SSD) [20], and sum of absolute differences (SAD) [18]. These common measures can be solved efficiently and effectively using concise optimizers such as classical gradient descent or brute force matching.

In this work, we propose the phase congruency network (PCNet) as a similarity enhancement method for multispectral/multimodal image registration. PCNet is a trainable network constructed under the guidance of phase congruency theory. Phase congruency theory [21,22] is a method of feature perception in the images which is invariant under illumination and contrast change. It has been employed in several tasks including similarity assessment [23,24], feature detection/extraction [25,26], image fusion [27,28], and image registration [29–31].

PCNet detects edge structures by estimating the phase congruency of different scales of frequency components. The magnitude of the output feature maps of PCNet merely depends on the congruency of phase among different frequency scales. Therefore, the similarity of the output features will be significantly enhanced in spite of the nonlinear variation of intensity and gradient. According to our experiments, PCNet outperforms state-of-the-art similarity enhancement methods and feature-based methods. As a trainable architecture, PCNet owns the advantage of the injection of the prior knowledge from the phase congruency theory together with the registration framework. The trainable parameters of PCNet is 2400× and 1500× less than the deep-learning registration method deep homography network (DHN) [32] and unsupervised deep homography network (UDHN)[1] [33], while the registration performance of PCNet surpasses them. What is more, PCNet can act as a complementary part of the deep-learning registration methods such as DHN, multiscale homography network (MHN) [34] and UDHN [33], which can significantly boost their registration accuracy. For example, the percentage of the number of images under 1 pixel average corner error (ACE) of UDHN is raised from 0.1% to 82.5% after the processing of PCNet, on CAVE dataset. It is worth noting that as is explored in [34], a more combination of the CNN architectures cannot guarantee such accuracy.

The multispectral/multimodal image registration procedure using PCNet is performed as follows. First, we build the phase congruency architecture under the guidance of phase congruency theory, which contains 2 trainable layers. Second, we employ the modified learnable Gabor kernels for multi-scale frequency component extraction, which significantly reduce the difficulty of network training and improve the generalization ability at the same time. Third, we introduce a normalized structural similarity loss for the no ground truth training of PCNet. Finally, we present the image registration framework using the image pyramid and gradient descent algorithm.

To summarize, the main contributions of this work are as follows:

- Based on phase congruency, we propose the phase congruency network (PCNet) that can significantly improve the structural similarity between images having nonlinear variation of intensity and gradient. The network contains two parts, i.e., the phase congruency architecture and modified learnable Gabor kernels.
- To cope with the problem that no ground truth exists for phase congruency features, we design a normalized structural similarity loss depicting the similarity between the output feature maps of PCNet. The network can then be trained without the need of the ground truth.
- We show that thanks to the prior knowledge of phase congruency, PCNet can be trained once on multispectral data but works effectively on a wide range of multispectral/multimodal data. PCNet can not only outperform state-of-the-art registration methods but also works as a complementary part of the deep-learning methods which significantly boost their registration accuracy.

2. Related work

The registration techniques for multispectral/multimodal images can be coarsely categorized into intensity-based methods and feature-based methods. In the following, we will present a brief review of the similarity enhancement intensity-based methods. For more detailed surveys on image registration, please refer to [16,35].

For intensity-based methods, we focus on the similarity enhancement ones as they are generally more efficient and easy to optimize compared to the conventional complicated multispectral/multimodal measures such as mutual information (MI) [36], robust selective normalized cross correlation (RSNCC) [13], and residual correlation ratio (CR) [37].

Entropy image (EI) [20] obtains the consistent structure using the Shannon entropy. EI adopts local histogram to compute the probability of different intensity levels and then computes the entropy as the enhanced structure.

Weber local descriptor (WLD) [38] detects local texture information based on Weber's law [39]. The law states that for the human visual system, the perceived change in stimuli is proportional to the initial stimuli. In [40], the differential excitation component of WLD is employed for multimodal image registration.

Structure consistency boosting (SCB) transform [18] adopts the statistical prior from natural images. The consistent inherent edge structures are enhanced using a transform depicted by generalized Gaussian distribution (GGD) with 3 trainable parameters.

Dense adaptive self-correlation (DASC) [41] descriptor is also a similarity enhancement algorithm leveraging machine learning. The algorithm is established based on the local self-similarity prior. The similarity between patches in a local support window is computed and encoded into a descriptor. The optimal sampling pattern for patch similarity computation is learned using the support vector machine (SVM).

Several feature-based methods have been proposed for registering multispectral/multimodal images. Log-Gabor histogram descriptor (LGHD) [42] and radiation-variation insensitive feature transform (RIFT) [31] employ the log-Gabor filter to alleviate the nonlinear variation of intensity and gradient. LGHD builds a histogram for the filter response of each scale of the log-Gabor filter. The combined histogram of different scales finally forms a descriptor. Different from LGHD, RIFT does the summation over each scale of the log-Gabor filter and then builds a maximum index map (MIM) by recording the channel number of the maximum value through different orientations of the log-Gabor filter responses. A histogram similar to Scale-invariant feature transform (SIFT) [43] is established based on the MIM and is used as the final descriptor. On the other hand, efficient feature matching and position matching algorithm (MatchosNet) [44] employs a complex deep convolutional neural network consisting of multiple dense convolutional blocks and cross stage partial networks to generate feature descriptors that are robust under nonlinear variation of intensity and gradient.

Recently, many deep-learning methods have been proposed to directly register the unaligned images. Deep homography network (DHN) [32] adopts a VGG-style (the network architecture proposed by Visual Geometry Group) network [45] to directly predict the global motion between the concatenated source and target images. Multiscale homography network (MHN) [34] is then proposed to align image pairs by cascading 3 levels of VGG-style networks, which significantly improves the registration accuracy. Similarly, multiscale framework

[1] UDHN here denotes a designed network architecture, and is trained in the supervised manner for fair comparison.

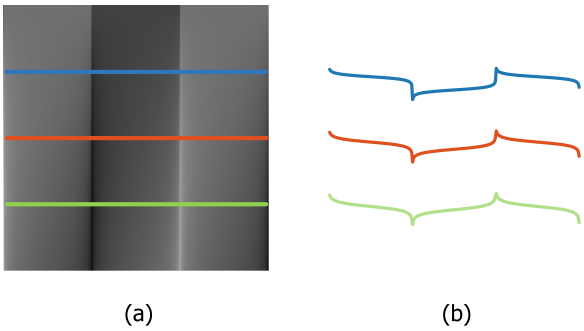


Fig. 1. The constructed grating and its corresponding profiles. (a) the grating constructed using (1). (b) the corresponding profiles from (a).

with unsupervised learning (MU-Net) [46] registers the remote sensing images with the stacked deep neural network models on multiple scales. MU-Net is trained in an unsupervised manner under the loss function constructed by channel feature of orientated gradient (CFOG) [47]. Another network based on ResNet-34 [48] is also proposed to align images with moving foregrounds, named unsupervised deep homography network (UDHN) [33].

3. Phase congruency theory

In an image, edge features are perceived where the different scales of wavelets meet maximum phase congruency. For better illustration, we construct a grating in Fig. 1(a) using the series

$$\sum_{s=0}^{\infty} \frac{1}{2s+1} \sin((2s+1)x + \phi), \tag{1}$$

where x varies along the horizontal direction, and ϕ the vertical, denoting the congruent phase shift. s denotes the scale of the wavelet series. Profiles of different phase shifts are drawn in Fig. 1(b). We observe that the congruency of phase at every phase shift produces a clearly perceived feature, namely the edge structure of an image.

Different practical ways for calculating phase congruency have been proposed in [21,22,49,50], and a robust approach was introduced in [22]. Concretely, the phase congruency feature is computed by first convolving the image with a quadrature pair of wavelet filters at each scale s. The wavelet filters contain the even-symmetric and odd-symmetric parts. The frequency responses of the wavelet filters at each scale are denoted by $e_s(x, y)$ for the even and $o_s(x, y)$ the odd. We can then compute the amplitude of response at each scale

$$A_s(x, y) = \sqrt{e_s(x, y)^2 + o_s(x, y)^2}, \tag{2}$$

and the phase

$$\phi_s(x, y) = \arctan(o_s(x, y)/e_s(x, y)). \tag{3}$$

Then the local energy can be computed as

$$E(x, y) = \sqrt{(\sum_s e_s(x, y))^2 + (\sum_s o_s(x, y))^2}, \tag{4}$$

and its corresponding phase as

$$\bar{\phi}(x, y) = \arctan(\sum_s o_s(x, y) / \sum_s e_s(x, y)). \tag{5}$$

For clarity, we refer to the phase of multi-scale filter response as phase (denoted by $\phi_s(x, y)$), and the phase of local energy as mean phase (denoted by $\bar{\phi}(x, y)$) in the following.

The phase congruency at pixel position (x, y) is measured by the ratio of the local energy and the accumulation over scales of the amplitude of response

$$\begin{aligned} PC_0(x, y) &= \frac{E(x, y)}{\sum_s A_s(x, y) + \xi} \\ &= \frac{\sum_s A_s(x, y) \cos(\phi_s(x, y) - \bar{\phi}(x, y))}{\sum_s A_s(x, y) + \xi}, \end{aligned} \tag{6}$$

where ξ is a small value that avoids division by zero. It is claimed that the above phase deviation estimation calculation is prone to produce blurry phase congruency features [22]. To endow the phase deviation estimation with more ocular discriminability, a correction term is added in. The phase congruency is then formulated as

$$PC_1(x, y) = \frac{\sum_s A_s(\cos(\phi_s - \bar{\phi}) - |\sin(\phi_s - \bar{\phi})|)}{\sum_s A_s + \xi}, \tag{7}$$

where the pixel position (x, y) is omitted for notation simplification.

The features extracted by the phase congruency procedure can weaken the inconsistency of the original image edge. As a similarity enhancement technique, phase congruency theory has been adopted in many multispectral/multimodal image registration methods, such as automatic registration of remote-sensing images (ARRSI) [29], histogram of orientated phase congruency (HOPC) [30], and the above-mentioned RIFT [31]. ARRSI detects feature points and performs normalized cross-correlation (NCC) [51] matching on the maximum moment map of phase congruency. HOPC combines the orientation and the amplitude of phase congruency with the histograms of oriented gradients (HOG) descriptor to construct a similarity enhanced feature descriptor. RIFT employs the maximum moment map of phase congruency to conduct feature detection. Different from the above registration methods that directly use the product of phase congruency computation [22], PCNet intends to achieve the implementation of the phase congruency theory from a completely new perspective, which is a learnable manner that can be trained to be more suitable for the desired task. The concept of implementing the phase congruency theory in a learnable manner is novel and might be further adopted into other phase-congruency-based methods such as [30,31,52–55].

4. Phase congruency network for structural similarity enhancement

For multispectral and multimodal images, the phase congruency procedure can enhance structure consistency regardless of the nonlinear variation of image intensity and gradient. However, the present phase congruency procedure is oversensitive to image noise and fake edges. To produce satisfactory similarity enhancement results, we proposed the phase congruency network (PCNet).

Our PCNet is constructed based on phase congruency [22]. Several modification strategies are put forward for network construction. Fig. 2 depicts the schematic diagram of PCNet. The input images are first convoluted by the modified learnable Gabor kernels and then fed into the phase congruency architecture. The similarity enhanced structure outputs are finally produced by PCNet. Note that Fig. 2 illustrates the inference stage of PCNet, in which the two input images are not aligned. For the training stage please kindly refer to Fig. 8. The modified learnable Gabor kernels consist of the learnable convolutional kernels and the fixed Gabor wavelets. The Gabor wavelets contain quadrature pairs of wavelets, and are steerable and scalable, satisfying the phase congruency theory. The phase congruency architecture is constructed based on the aforementioned theory with two trainable layers, namely noise estimation layer and modified phase deviation estimation layer. We will describe the details of PCNet below.

We first reformulate (7) into a more compact form for tensor manipulation,

$$\mathbf{P} = (\sum_s \mathbf{A}_s \circ \Delta\mathbf{\Phi}_s) \oslash (\sum_s \mathbf{A}_s + \xi \cdot \mathbf{1}), \tag{8}$$

where $\oslash$ denotes pointwise division or Hadamard division [56]. $\mathbf{1} \in \mathbb{R}^{H \times W}$ is an all-ones matrix. $\xi \cdot \mathbf{1}$ prevents the numerator from being

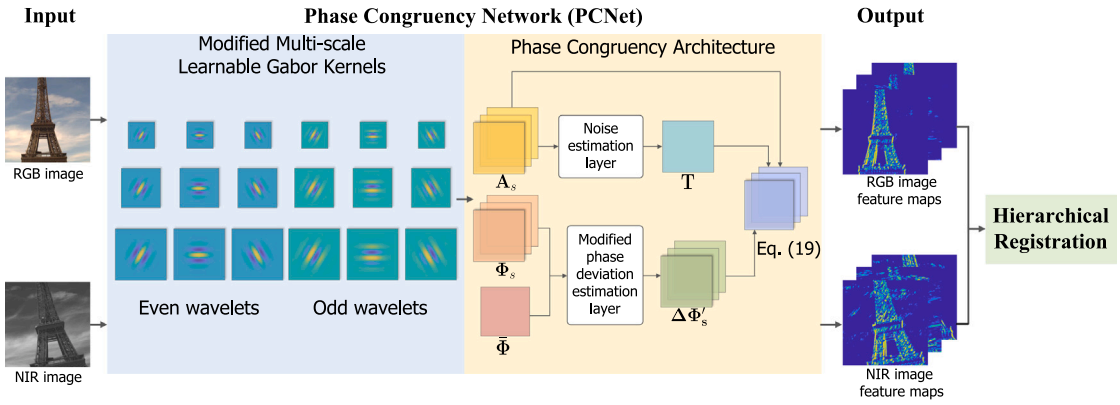


Fig. 2. The schematic diagram of phase congruency network (PCNet). The network is constructed based the phase congruency. Two trainable layers (noise estimation layer and modified phase deviation estimation layer) are learnable parts of the phase congruency architecture. The modified learnable Gabor kernels are employed as the quadrature wavelet bank satisfying phase congruency theory. The input RGB and NIR images are processed by PCNet and then transformed into similarity-enhanced feature maps. The feature maps are then fed into the hierarchical registration module. Note that, for illustration, the diagram shows the phase congruency architecture within one orientation, and shows the similarity enhancement in the inference stage.

divided by zero and will be omitted in the following. $\mathbf{P} \in \mathbb{R}^{H\times W}$, $\mathbf{A}_s \in \mathbb{R}^{H\times W\times S}$, and $\Delta\mathbf{\Phi}_s \in \mathbb{R}^{H\times W\times S}$, with S being the amount of filter scales. $\Delta\mathbf{\Phi}_s$ indicates the phase deviation estimation layer, defined as

$$\Delta\mathbf{\Phi}_s = \cos(\mathbf{\Phi}_s - \bar{\mathbf{\Phi}}) - |\sin(\mathbf{\Phi}_s - \bar{\mathbf{\Phi}})|, \tag{9}$$

where $\mathbf{\Phi}_s$ denotes the phase map and $\bar{\mathbf{\Phi}}$ the mean phase map.

To obtain satisfactory similarity enhancement results for image registration, two trainable layers are proposed and then combined into PCNet.

4.1. Noise estimation layer

The computation of phase congruency leveraging (8) is sensitive to noise because in the natural image noise forms small edges. Hence, the noise should be estimated and eliminated before computing the phase congruency. As illustrated previously, the local energy is the square root of two independent random variables, each following a standard normal distribution. Thus the noise of the local energy will have a Rayleigh distribution. We denote $\mathbf{M}_\mathrm{R}$ as the mean of the Rayleigh distribution as

$$\mathbf{M}_\mathrm{R} = \sqrt{\frac{\pi}{2}} \cdot \mathbf{V}_\mathrm{G}, \tag{10}$$

and $\mathbf{V}_\mathrm{R}$ as the variance of the Rayleigh distribution as

$$\mathbf{V}_\mathrm{R} = \sqrt{\frac{4-\pi}{2}} \cdot \mathbf{V}_\mathrm{G}. \tag{11}$$

The Rayleigh noise map can then be estimated as

$$\mathbf{T} = \mathbf{M}_\mathrm{R} + \mathbf{V}_\mathrm{R}. \tag{12}$$

$\mathbf{V}_\mathrm{G}$ in (10) and (11) denotes the variance of the end position of the local energy vector, which is estimated as

$$\mathbf{V}_\mathrm{G} = \tau \cdot (\mathbf{1} - (\mathbf{1}/\alpha)^{\circ N_s}) \oslash (\mathbf{1} - \mathbf{1}/\alpha + \xi \cdot \mathbf{1}), \tag{13}$$

where N_s denotes the number of the frequency scales and $^\circ$ denotes the pointwise power or Hadamard root [57]. τ can be directly estimated from the local energy. Note that the original α here is defined as the scaling factor between successive filters, but in this work, we make it a trainable unit as it directly controls the overall noise threshold. Parameter α will be updated by the gradient flow in the training stage, making the Rayleigh noise map into the noise estimation layer of our PCNet.

Once having the threshold produced by the noise estimation layer, the noise can be removed by soft thresholding

$$f(m) = \begin{cases} m - t & \text{if } m > t \\ 0 & \text{if } m \le t \end{cases}, \tag{14}$$

where t denotes the estimated threshold. For our PCNet, the soft thresholding operation can be perfectly modified into the rectified linear units [58], namely the ReLU activation function, which can then be formulated as

$$\mathbf{P} = \mathrm{ReLU}(\sum_s \mathbf{A}_s \circ \Delta\mathbf{\Phi}_s - \mathbf{T}) \oslash \sum_s \mathbf{A}_s. \tag{15}$$

4.2. Modified phase deviation estimation layer

The original phase deviation estimation layer according to [22] is formed by (9). The added correction term can make the output phase congruency features visually thinner. However, it is not clear that ocular discriminability is beneficial to structure similarity enhancement. Hence, we employ a trainable term β in the phase deviation estimation layer, yielding the modified phase deviation estimation layer

$$\Delta\mathbf{\Phi}'_s = \cos(\mathbf{\Phi}_s - \bar{\mathbf{\Phi}}) - \beta \cdot |\sin(\mathbf{\Phi}_s - \bar{\mathbf{\Phi}})|. \tag{16}$$

By adopting the above trainable layers, our PCNet has the mathematical form

$$\mathbf{P}(\alpha, \beta) = \mathrm{ReLU}(\sum_s \mathbf{A}_s \circ \Delta\mathbf{\Phi}'_s - \mathbf{T}) \oslash \sum_s \mathbf{A}_s. \tag{17}$$

With the assistance of phase congruency, we have constructed a network architecture with trainable units. However, we still need a proper set of wavelet filters for multi-scale frequency component extraction.

4.3. Modified multi-scale learnable gabor kernels

It is a straightforward idea to construct a series of convolutional kernels for multi-scale frequency component extraction. However, there are no ground truth phase congruency feature maps for network training, and hence it is difficult to train the convolutional kernels without regularization. Furthermore, the large-scale frequency components of phase congruency generally require relatively huge convolutional kernel sizes (e.g. a kernel size of 25 × 25), which also increases the difficulty of kernel training.

Considering the above issues, we adopt modified multi-scale learnable Gabor kernels as part of the network architecture for multi-scale frequency component extraction. As illustrated in Fig. 3, the Gabor wavelets are steerable and scalable filters, which is created by Dennis Gabor [59]. It is claimed that simple cells in the visual cortex of mammalian brains can be modeled by the Gabor functions [60], thus the Gabor wavelets are thought to be similar to the perception of the human visual system. What is more, it has been shown that the shallow layers of image-trained convolutional neural networks (CNNs) tend to learn filters resembling Gabor filters [61]. Gabor filters are composed

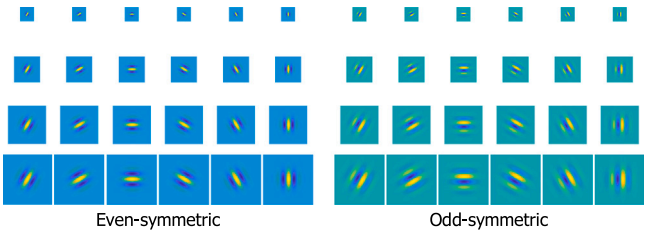


Fig. 3. Examples of the steerable and scalable Gabor wavelets.

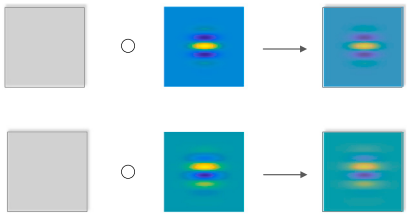


Fig. 4. Illustration of the learnable Gabor filters.

of a pairwise bank of multi-scale quadrature wavelets, which perfectly satisfies the requirement of the phase congruency theory. The Gabor wavelets are defined as

$$G_{u,v}(\mathbf{z}) = \frac{\|\mathbf{k}_{u,v}\|^2}{\sigma^2} e^{\left(-\|\mathbf{k}_{u,v}\|^2\|\mathbf{z}\|^2/2\sigma^2\right)} \left(e^{i\mathbf{k}_{u,v}^{\mathrm{T}}\mathbf{z}} - e^{-\sigma^2/2}\right). \tag{18}$$

The Gabor filters are made learnable by making pointwise production with the learnable convolutional kernels having the same size [61] as illustrated in Fig. 4. The original purpose of introducing Gabor wavelets into CNN architecture in [61] is to guide the learnable convolutional kernels with directional information, which lightens the image classification deep networks and improves classification performance. On the contrary, in our work, the modulation provides the CNN layer with more constraints, which significantly reduces the difficulty of network training. It is worth noting that compared to [61], our modified learnable Gabor kernels are markedly different in 3 aspects:

- We employ the full parts (even-symmetric and odd-symmetric) of Gabor wavelets in accordance with the phase congruency theory, while [61] only uses the even-symmetric part to guide the direction of convolutional kernels for the classification purpose.
- We achieve the multi-scale convolution by directly enlarging the size of kernels at different scales, which perfectly matches the requirement of phase congruency theory, whereas [61] resizes the input features using the max-pooling operation.
- Considering that the Gabor filters have the drawback of being over-sensitive to the direct current (DC) component of the signal, we modify the Gabor wavelets by subtracting their corresponding averages.

To illustrate the effectiveness of the modified Gabor filters, we draw the filter responses together with the phase congruency outputs for the original Gabor filters and the modified ones in Fig. 5. It is observed that without the modification, the filter has an obvious response for the DC components of the image. As a result, the corresponding phase congruency output is also sensitive to the DC component. On the contrary, our modified Gabor filters significantly alleviate this problem.

For each orientation of the modified learnable Gabor kernels, the phase congruency architecture adopts the corresponding filter outputs and produces a channel of the phase congruency map. At last, channels of map form multi-channel phase congruency feature maps. The dimension of the features is determined by the orientation number of the modified learnable Gabor kernels. By denoting o as the orientation,

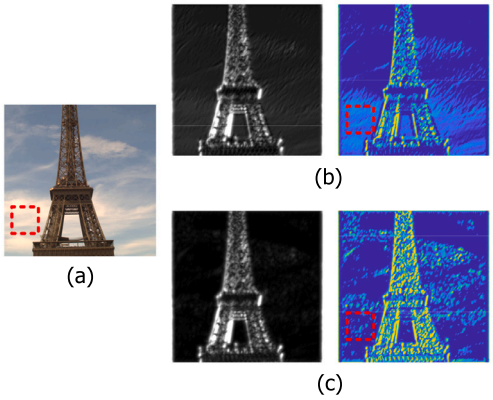


Fig. 5. The filter responses together with the phase congruency outputs for (b) the original Gabor filters and (c) the modified Gabor filters. The red boxes highlight the areas for detailed comparison.

the calculation of our PCNet is finally formulated as

$$\mathbf{P}_o(\alpha, \beta, W) = \mathrm{ReLU}\left(\sum_s \mathbf{A}_{s,o} \circ \Delta\Phi'_{s,o} - \mathbf{T}_o\right) \oslash \sum_s \mathbf{A}_{s,o}, \tag{19}$$

where W denotes the trainable parameters in the modified learnable Gabor kernels.

4.4. Normalized structural similarity loss for no ground truth network training

Our PCNet can be trained by the stochastic gradient descent (SGD) algorithm using a proper loss function. However, as previously mentioned, the ground truth phase congruency feature maps do not exist for network training. In this work, we employ a loss function that directly compares the pairwise similarity of the output phase congruency feature maps of two input multispectral or multimodal images. In this way, a no ground truth learning framework is established, which perfectly matches our requirement of enhancing the structural similarity of the input images.

The network is trained using the siamese strategy. Let I_1 and I_2 be two original band images, $\mathbf{P}_1$ and $\mathbf{P}_2$ be the outputs after applying our PCNet. Considering that the phase congruency features produced by PCNet are similar to edge structures, we employ structural similarity index measure (SSIM) [62] to measure the similarity of $\mathbf{P}_1$ and $\mathbf{P}_2$,

$$L = 1 - \sum_o \mathrm{SSIM}(\mathbf{P}_{1,o}, \mathbf{P}_{2,o})/N_o. \tag{20}$$

Here comes a problem that the above loss function tends to produce an ambiguous guide for the network training. For example, a pair of outputs with all 0 intensity will produce a minimized loss function with 0 value. Consequently, the structural content of the image will be exterminated and the registration procedure will fail. To cope with this problem, we adopt the gradient of an image to protect its structural information, yielding the modified loss function

$$L = \frac{1 - \sum_o \mathrm{SSIM}(\mathbf{P}_{1,o}, \mathbf{P}_{2,o})/N_o}{|\sum_l \sum_o (\|\nabla_l \mathbf{P}_{1,o}\|_1 + \|\nabla_l \mathbf{P}_{2,o}\|_1)/N_o|^c}, \tag{21}$$

where the operator ∇_l, $l \in \{x, y\}$, represents the gradient computation along the horizontal and vertical directions. c is the structure protection parameter that balances the degree of similarity enhancement and structure protection. The networks can be trained using the SGD algorithm.

4.5. Hierarchical registration framework

We adopt a hierarchical intensity-based framework to accomplish the image registration, which optimizes a predefined similarity measure

with reference to the parametric or non-parametric transformations. In this work, we employ the parametric affine transformation model that can handle image deformation such as rotation, scaling, translation, shearing and any combinations of them [63]. The model has been widely adopted for multispectral and multimodal image registration [14,64], which is formulated by

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x & y & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & y & 1 \end{pmatrix} \mathbf{a}, \tag{22}$$

where $p = (x, y)^{\mathsf{T}}$ and $\tilde{p} = (u, v)^{\mathsf{T}}$ denote, respectively, the pixel coordinates before and after affine transformation, and $\mathbf{a} = (a_1, a_2, a_3, a_4, a_5, a_6)^{\mathsf{T}}$ denotes the affine transformation parameters, which are optimized in the framework.

As for the similarity measure, we use the sum of squared differences (SSD). SSD is efficient and easy to optimize, but sensitive to inconsistencies in image brightness in multispectral/multimodal images. Fortunately, the inconsistency of the input images is significantly weakened by our proposed PCNet, which makes efficient and effective registration using SSD possible. The SSD measure is formulated as

$$J(\mathbf{a}) = \sum_{p \in \Omega(\mathbf{a})} (F_{\mathrm{R}}(p) - F_{\mathrm{F}}(p, \mathbf{a}))^2, \tag{23}$$

where F_{R} and F_{F} denote the reference and floating feature maps, and $\Omega(\mathbf{a})$ denotes the meaningful overlapping area of the warped I_F. And hence the objective function of the optimization framework is

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}}\, J(\mathbf{a}), \tag{24}$$

where $\hat{\mathbf{a}}$ denotes the final optimal affine parameters. We note that for the feature map produced by PCNet with multiple channels, SSD is computed on all the channels and then summed, which can be formulated as

$$J_{\mathrm{PCNet}}(\mathbf{a}) = \sum_{o} \sum_{p \in \Omega(\mathbf{a})} (\mathbf{P}_{o,\mathrm{R}}(p) - \mathbf{P}_{o,\mathrm{F}}(p, \mathbf{a}))^2. \tag{25}$$

To obtain the optimal registration parameters $\hat{\mathbf{a}}$, we employ the hierarchical registration based on the Gaussian pyramid, which is illustrated in Fig. 6. The feature maps from PCNet are continually downsampled by a scale factor of s with Gaussian blur until the spatial size of the feature map reaches a threshold. The feature pyramid of N scales is then constructed. In the optimization stage, the constructed feature pyramid works from the bottom to the top, which is also called the coarse-to-fine manner. The affine parameters are first initialized as $\mathbf{a}^0 = (1, 0, 0, 0, 1, 0)^{\mathsf{T}}$, and then updated in feature level l_1 to produce $\mathbf{a}_1$. $\mathbf{a}_1$ is then transferred by multiplying the scale factor s into the translation components to facilitate the parameter update in the next level l_2. The interleaved transfer (initialization)-update process works successively on each level of the feature pyramid until the final level l_N, and produces the final estimated $\hat{\mathbf{a}} = \mathbf{a}_N$. In our implementation, we empirically set the scale factor $s = 2$, and the threshold of the spatial size as 16×16.

In each feature level, the registration parameters are updated using the gradient descent optimization

$$\mathbf{a}^{t+1} = \mathbf{a}^t - \eta \nabla_{\mathbf{a}} J(F_{\mathrm{R}}, F_{\mathrm{F}}, \mathbf{a}^t), \tag{26}$$

where η denotes the step size, t the iteration number during the optimization, and $\nabla_{\mathbf{a}} J(F_{\mathrm{R}}, F_{\mathrm{F}}, \mathbf{a}^t)$ the gradient of (23) with respect to $\mathbf{a}^t$. The iteration stops when it either reaches the maximum iteration number of each level T_n or the update components are smaller than a threshold $\epsilon = 0.0000000001$. We set η in each level as $0.75/\max(H_n, W_n)$, where H_n and W_n denote the spatial size of the feature map. And we set the initial maximum iteration number of level l_1 as $T_1 = 500$, which is decreased by half for efficiency every time the optimization gets into the next feature level.

We note that the above details together with the hierarchical image registration framework are the same for all the experiments for PCNet and the compared similarity enhancement methods, which guarantees fair evaluation and comparison.

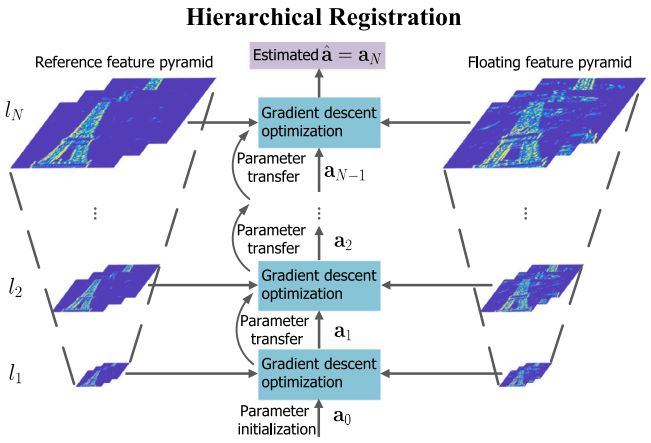


Fig. 6. The hierarchical registration process leveraging Gaussian pyramid and output feature maps from PCNet.

4.6. Discussion about the difference of loss function for PCNet and objective function for iterative registration

The process that producing phase congruency feature maps using PCNet can be roughly viewed as an image-to-image translation task, and hence point-wise supervision like SSD is unable to describe the similarity of the output feature maps. For example, as illustrated in [62], SSD (namely MSE in [62]) cannot depict image inconsistency under JPEG compression, blur, and salt–pepper impulsive noise contamination, etc. On the contrary, SSIM is composed of the patch-wise similarity description of the image luminance, contrast, and structure. The patch-wise similarities are then averaged to produce the description of the whole image, producing a more powerful measure. [65] also reveals the fact that using multiscale SSIM loss to depict the similarity of the CNN restored image and ground truth image performs a lot better than SSD (namely ℓ_2 in [65]). As for the objective function for iterative registration, the problem comes into a totally opposite way. We employ SSD for image registration because it is robust under the image distortions such as blur and noise. SSD also owns the advantage in efficiency because it is calculated in a point-wise way. The detailed comparison of registration accuracy using different loss functions for PCNet is shown in Section 5.1.

5. Experiments

Our PCNet is trained once on a subset of the CAVE [3] multispectral dataset, and evaluated on various multispectral/multimodal datasets including the CAVE and Harvard [4] multispectral band image datasets, visible/near-infrared (RGB/NIR) [1] pairwise multispectral image dataset, and flash/no-flash [2] dataset. The training subset for our PCNet is not included in the performance evaluation. We illustrate the above-mentioned datasets in Figs. 9 and 12. The scenes employed for further illustration are marked with *S1* ∼ *S7*.

The registration performance of our PCNet is compared with other state-of-the-art similarity enhancement algorithms including entropy image (EI) [20], WLD [38], and structure consistency boosting (SCB) transform [18]. The same aforementioned hierarchical registration strategy is employed for all the above similarity enhancement algorithms. We also compare our registration framework using PCNet with state-of-the-art feature-based multispectral/multimodal methods including log-Gabor histogram descriptor (LGHD) [42], radiation-variation insensitive feature transform (RIFT) [31], and dense adaptive self-correlation (DASC) descriptor [41]. As DASC produces dense 128-channel output similarity enhancement feature maps, we employ the

registration strategy provided by its public source code[2], which registers the feature maps using SIFT flow [66]. The output flow is then fed into the estimateGeometricTransform [67] function in Matlab, producing the affine registration result for a fair comparison. The affine transformation of LGHD and RIFT is also obtained by the estimateGeometricTransform function. For a better comparison, we adopt the registration error as the average Euclidean error (AEE) [14,18] between the pixel positions computed by the ground truth transformation $\mathbf{a}_{\text{gt}}$ and the estimated transformation $\hat{\mathbf{a}}$,

$$e_1 = \frac{1}{M}\sum_{p=1}^{M} \|\tilde{p}(p, \mathbf{a}_{\text{gt}}) - \tilde{p}(p, \hat{\mathbf{a}})\|_2, \tag{27}$$

where M indicates the number of pixels of an image, and $\tilde{p}$ the warped pixel position by applying the transformation $\mathbf{a}$ on pixel p.

For an extensive evaluation, we keep the reference image fixed and warp the floating image with the completely random affine transformations with the largest rotation up to 30°, scale up to 1.3, shearing up to 0.3, and center point translation up to 40 pixels. The process of random transformation is repeated 3 times for a data set to enlarge the test sample. In the comparison with the above methods, the above deformations are evaluated on 256 × 256 images if not otherwise specified.

We further conduct a comparison of our registration framework with the deep-learning registration methods including DHN [32], MHN [34], and UDHN [33]. We also investigate the performance improvement by combining our PCNet with the above deep-learning methods as in [68]. As the deep-learning networks are constructed to produce the displacement of the corner points of an image, the average corner error (ACE) [32] is used to conduct the registration performance comparison. ACE is formulated as

$$e_2 = \frac{1}{4}\sum_{p=1}^{4} \|\tilde{p}(p, \mathbf{a}_{\text{gt}}) - \tilde{p}(p, \hat{\mathbf{a}})\|_2, \tag{28}$$

where the main difference compared to AEE is the error is only calculated on the 4 corner points of an image instead of all points. The size of the image to be registered is reduced to 128 × 128 following the design of the deep-learning networks. The translation parameters of the simulated transformation are reduced in proportion simultaneously.

We employ two ways to display the registration error. The first one is the table including the mean, median, trimean,[3] and the mean of the errors below 25th, 50th, 75th, and 95th percentiles (denoted by best25%, best50%, best75%, and best95%) within a dataset as in [18]. Another one is the figure that plots the fraction of the number of images with respect to the registration error as in [34].

In the following, we first discuss some issues of our PCNet, including the training details and the hyperparameter settings. We then compare our PCNet with the original phase congruency algorithm. Next, we compare our PCNet with the aforementioned registration methods including the conventional ones and deep-learning ones. We also analyze the number of parameters of our PCNet and other deep-learning methods, together with the performance improvement by combining our PCNet with them. Finally, we conduct an ablation study of our PCNet.

5.1. Training details and hyperparameter setting

Our PCNet is trained by a subset of the CAVE [3] dataset. The dataset consists of 32 multispectral image scenes, with each including 31 band images ranging from 400 to 700 nm. Example images from the CAVE dataset are displayed in Fig. 7(a). The first 10 scenes in alphabetical order are taken as the training data and the rest as test data. Specifically, we randomly crop several pairs of patches with the size of 200 × 200 for network training. The patches are fed into the network in pair and random order, with data augmentation (brightness, contrast, and saturation adjustment). The training process of PCNet is illustrated in Fig. 8. The input pairwise patches are fed into two identical PCNets of shared weights, and then the normalized structural similarity of output features are evaluated using the training loss Eq. (21). The weights of PCNet are updated by the gradient flow produced by the SGD optimizer. Note that our PCNet can be regarded as fully convolutional, which means that the network to work on any size of input image regardless of the patch size for the network training.

We set the scales for the phase congruency estimation as 4, with the scaling factor between successive filters being 2. For the 4 scales of the modified learnable Gabor kernels, the sizes of the learnable convolutional kernels are set as 7 × 7, 13 × 13, 19 × 19, and 25 × 25. It is worth noting that although the above learnable kernels seem to be overly large for common CNNs, they function well under the regularization of Gabor wavelets. The modified learnable Gabor kernels are set to stride 1. For the two trainable layers, we set the initial value of $\alpha = 2$ as it is originally defined as the scaling factor, and we set $\beta = 1$. As for other training details, we adopt the SGD optimizer with an initial learning rate of 0.01 and weight decay of 0.00003. The network training finishes after 25 epochs with a batch size of 100.

We compare the registration performance by training PCNet using the switched loss functions based on sum of the squared differences (SSD), sum of the absolute differences (SAD), and structural similarity index measure (SSIM). The gradient term for SSD is set to be quadratic and SAD absolute value. We note that all other settings are kept the same except the loss functions. The error statistics are illustrated in Table 1. It is observed that SSIM based loss indeed performs better.

We then evaluate two hyperparameters of PCNet, which are the structure protection parameter c in (21) and the orientation number of the modified learnable Gabor kernels N_o in (19). We evaluate the effect of both parameters in terms of registration accuracy. For the test data of CAVE, we take the 16th band image (i.e. 550 nm) in each scene as the reference image and generate floating images by imposing the aforementioned simulated deformations to all band images. In this manner, we get 2046 image pairs for registration experiments.

As elaborated previously, hyperparameter c controls the degree of protection for the structural information, and a larger c means better protection. Nevertheless, if c grows too large, the similarity of the output phase congruency features will be violated. Table 2 lists the error statistics of our PCNet under various hyperparameter c. We can observe that the table supports the above principle. The best registration performance is obtained at $c = 0.7$.

The hyperparameter N_o controls the orientation number of the modified learnable Gabor kernels, and thus determines the channel of the output feature maps. Smaller N_o can reduce channel number,

Table 1
Error statistics produced by training PCNet using the switched loss functions based on SSD, SAD, and SSIM on the CAVE dataset. For brevity, bestN% is denoted as B.N%, median as Med., and trimean as Tri. The best indicators are in bold.

Loss	Mean	Med.	Tri.	B.25%	B.50%	B.75%	B.95%
SSIM	**13.33**	**0.31**	**7.67**	**0.07**	**0.14**	**2.72**	**10.98**
SSD	19.26	3.52	11.74	0.09	0.29	9.26	17.15
SAD	17.77	0.62	9.92	0.09	0.19	7.43	15.59

Table 2
Error statistics produced by image registration using PCNet with various degrees of protection parameter for the structural information on the CAVE dataset. For brevity, bestN% is denoted as B.N%, median as Med., and trimean as Tri. The best indicators are in bold.

c	Mean	Med.	Tri.	B.25%	B.50%	B.75%	B.95%
0.6	**13.24**	0.35	7.80	0.08	0.16	**2.58**	**10.88**
0.7	13.33	**0.31**	**7.67**	**0.07**	**0.14**	2.72	10.98
0.8	13.48	0.33	7.73	**0.07**	**0.14**	2.92	11.14

[2] https://seungryong.github.io/DASC/

[3] trimean = $\frac{Q_1+2Q_2+Q3}{4}$, where $Q1$, $Q2$, and $Q3$ denotes the 1st, 2nd, and 3rd quartiles.

Fig. 7. Example multispectral images in the visible spectrum (displayed in RGB) from the (a) CAVE and (b) Harvard datasets. The scenes to be used for illustration are marked with *S1*, *S4*, and *S5*.

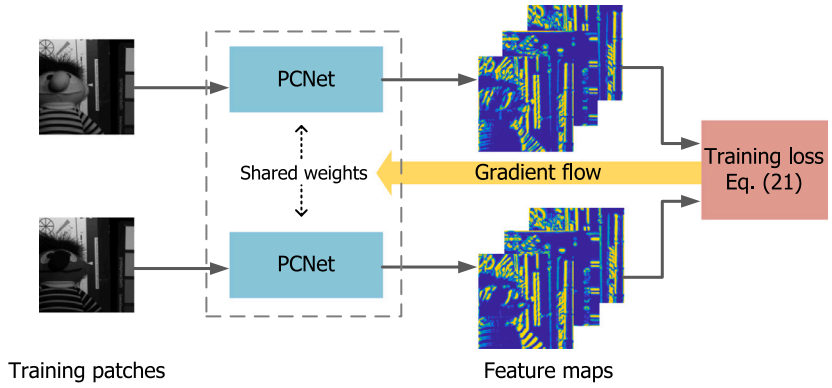


Fig. 8. The training process of PCNet. The aligned training patches are fed into two identical PCNets of shared weights. The output feature maps of PCNet are evaluated by the training loss, namely Eq. (21). The weights of PCNet are updated by the gradient flow produced by the stochastic gradient descent (SGD) optimizer.

Table 3
Error statistics produced by image registration using PCNet with various numbers of orientation of the modified learnable Gabor kernels on the CAVE dataset. For brevity, bestN% is denoted as B.N%, median as Med., and trimean as Tri. The best indicators are in bold.

N_o	Mean	Med.	Tri.	B.25%	B.50%	B.75%	B.95%
3	13.93	0.35	8.45	0.08	0.16	3.15	11.59
6	13.33	**0.31**	7.67	**0.07**	**0.14**	2.72	10.98
9	**13.05**	**0.31**	**7.60**	**0.07**	**0.14**	**2.46**	**10.70**

which can improve the efficiency of image registration with the side effect of worse accuracy. On the contrary, larger N_o is likely to produce better registration performance but the registration efficiency will be sacrificed. Table 3 lists the average errors of our PCNet under various hyperparameter N_o. It is observed that the best registration performance can be achieved at $N_o = 9$. However, the improvement of the registration accuracy from $N_o = 6$ to $N_o = 9$ is not significant compared to the cost in computation. Therefore, we set the channel number $N_o = 6$.

5.2. Comparison with conventional phase congruency algorithm

Thanks to the modification strategy, our PCNet can achieve better registration performance than the conventional phase congruency algorithm. We conduct similar registration experiments on the CAVE dataset as in Section 5.1. To conduct an exhaustive comparison for both algorithms, we evaluate their registration performance with various orientation number. We note that, a fewer feature channel means higher registration efficiency in the registration step.

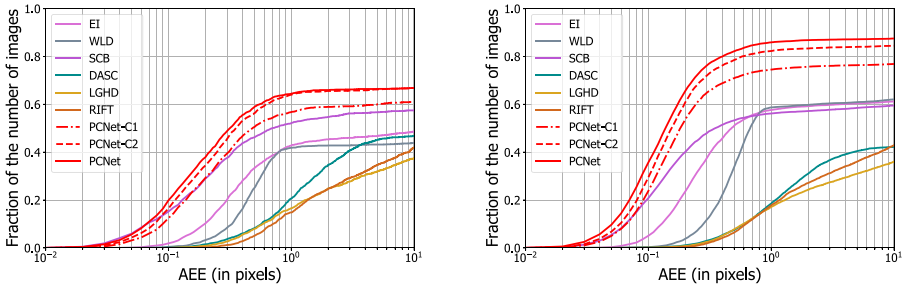
We list the error statistics for both algorithms of orientation number $N_o = 1$, $N_o = 2$, $N_o = 6$ and $N_o = 9$ in Table 4. The $N_o = 1$ PCNet and PC-org is obtained by the summation of the 2 orientations of $N_o = 2$ as it produces better results for both methods. The conventional phase congruency algorithm is denoted as PC-org. We can observe that within all feature channels, our PCNet produces considerably better results than PC-org. Furthermore, our PCNet with 2 channels could achieve higher registration accuracy than the PC-org with 9 channels, which means 4.5× improvement of computational efficiency in the registration step. Another interesting phenomenon occurs when we focus on the Best25% and Best95% statistics. For our PCNet, the Best25% and Best95% decrease as N_o increases, which means the overall accuracy improvement. On the contrary, the Best25% of PC-org increases as N_o increases, which means the loss of accuracy.

5.3. Results on multispectral band images

We evaluate our PCNet with other registration algorithms including EI [15], WLD [38], SCB [18], DASC [41], LGHD [42], RIFT [31] on

Table 4
Error statistics produced by image registration using PCNet and PC-org with various numbers of orientation of filters on the CAVE dataset. For brevity, bestN% is denoted as B.N%, median as Med., and trimean as Tri. The best indicators are in bold. Note that B.50% is omitted for a better article layout.

N_o	PCNet						PC-org					
	Mean	Med.	Tri.	B.25%	B.75%	B.95%	Mean	Med.	Tri.	B.25%	B.75%	B.95%
1	**15.66**	**0.48**	**8.95**	0.10	**5.10**	**13.39**	19.98	0.60	10.83	**0.07**	7.92	17.26
2	**13.51**	**0.35**	**7.90**	0.08	**2.90**	**11.16**	19.72	0.54	10.79	**0.06**	7.66	17.01
6	**13.33**	**0.31**	**7.67**	**0.07**	**2.72**	**10.98**	14.96	0.34	8.73	**0.07**	2.94	12.16
9	**13.05**	**0.31**	**7.60**	**0.07**	**2.46**	**10.70**	14.16	0.37	8.29	0.09	2.29	11.37



(a) Evaluation on CAVE dataset. (b) Evaluation on Harvard dataset.

Fig. 9. Registration evaluation on CAVE and Harvard datasets using PCNet and other registration algorithms. The fraction of the number of images within a dataset is plotted with respect to AEE.

the CAVE [3] and Harvard [4] datasets. The Harvard dataset contains 77 multispectral images of real-world scenes, each with 31 spectral bands ranging from 420 to 720 nm. The sample images are displayed in Fig. 7(b). Similar to the experiment setting for the CAVE dataset, we again take the 16th band image (i.e., 570 nm) of each Harvard scene as the reference image, and generate the floating images by imposing the simulated transformations. In this way, we conduct 2046 image pairs for image registration experiments on the CAVE dataset and 7176 image pairs on the Harvard dataset. We denote the 6 orientation version PCNet as PCNet, the 2 orientation version as PCNet-C2, and 1 orientation version as PCNet-C1. PCNet and PCNet-C2 are trained separately, and PCNet-C1 is obtained by the summation of the 2 orientations of PCNet-C2. For a better investigation of PCNet, if not specifically mentioned, we display the registration results of all the above versions of PCNet. We note that a fewer orientation number means a higher registration efficiency.

We plot the fraction of the number of images with respect to AEE within CAVE and Harvard datasets in Fig. 9. It is observed that all versions of PCNet enjoys the best registration performance on CAVE and Harvard dataset for most of the time, except for the slight performance degradation compared to SCB when AEE is lower than 0.2 pixels. As for other competitors, SCB and EI perform relatively better than WLD. As for the feature-based methods, they produce a considerable amount of registration AEE lower than 10 pixels. However, their registration result is not as accurate as PCNet.

To better investigate the registration performance, we further illustrate the registration results at different degrees of angular rotation and center point translation on CAVE and Harvard datasets in Fig. 10. We define **small** deformation as the angle of **rotation ranges from 0 to 10°**, and distance of **translation ranges from 0 to 20 pixels**; **middle** deformation as the angle of **rotation ranges from 10 to 20°**, and **distance of translation ranges from 10 to 30 pixels**; **large** deformation as the **angle of rotation ranges from 20 to 30°**, and distance of **translation ranges from 20 to 40 pixels**. It is observed that PCNet keeps producing the best registration performance under all degrees of deformation, with PCNet-C1 and PCNet-C2 outperforming other competitors in most cases. Most of the competitors can produce promising registration results under small deformation, while they encounter severe performance degradation as the degree of deformation grows. On the contrary, the advantage of PCNet over other methods becomes more obvious as the deformation becomes larger.

We further demonstrate the similarity enhanced outputs together with the corresponding SSD plots for all the similarity enhancement methods in Fig. 11. It is observed that the similarity enhanced outputs for EI have significant differences, which results in the SSD plot failing to indicate the best registration position. The similarity enhanced outputs of WLD and SCB are of weak consistency, and hence their SSD plots give weak guidance for the best registration position. In comparison, PCNet produces consistent similarity enhancement outputs, and its SSD plot is of a larger capture range and stronger minimum peak than any other algorithms.

5.4. Results on flash/no-flash and RGB/nir image pairs

We further evaluate our PCNet together with other registration methods on flash/no-flash and RGB/NIR datasets. The flash/no-flash dataset [2] includes 120 indoor and outdoor image pairs. The RGB/NIR dataset [1] contains 256 RGB/NIR image pairs of various categories of scenes. The example images from the above datasets are displayed in Fig. 12. For the flash/no-flash dataset, we set the flash image as the reference image and no-flash as the floating one. For the RGB/NIR dataset, we set the RGB image as the reference image and NIR as the floating one. The deformations imposed on the floating images are the same as in previous experiments. Totally we conduct 360 experiments on the flash/no-flash dataset and 768 experiments on the RGB/NIR dataset. It is worth noting that though our PCNet is trained on a subset of CAVE multispectral band data, experimental results show that it performs well on flash/no-flash and RGB/NIR datasets without any retraining.

We draw the fraction of the number of images with respect to AEE within flash/no-flash and RGB/NIR datasets in Fig. 13. It is observed that all versions of PCNet outperform other competitors without any retraining, which means a superior generalization ability for different image data. As for other competitors, they produce similar performance as in the previous Section 5.3.

Fig. 14 displays the registration results of scenes *S2* and *S3* produced by PCNet and other registration algorithms. The registration results are illustrated by overlapping the reference image and the registered

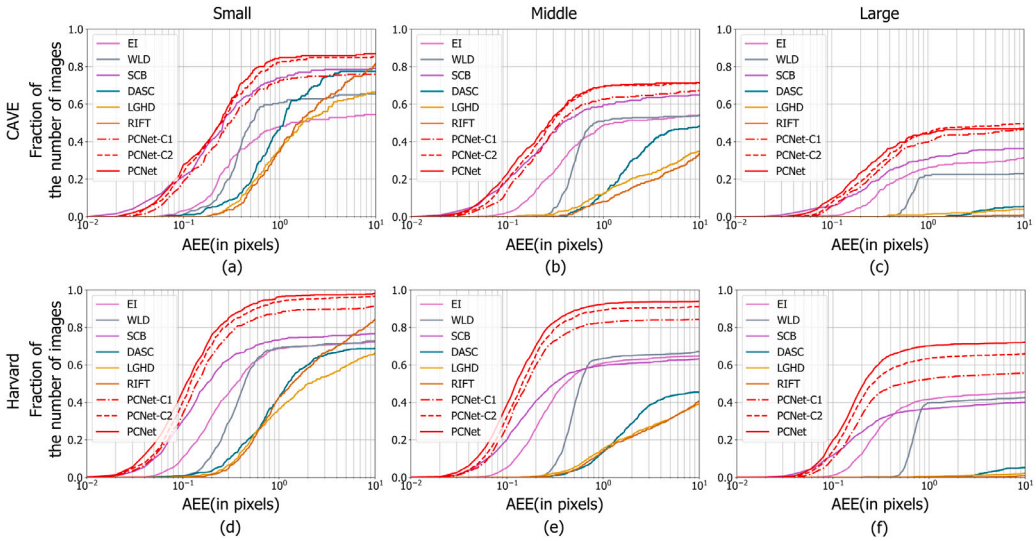


Fig. 10. Registration evaluation under small, middle, and large degree of transformations on CAVE and Harvard datasets using PCNet and other registration algorithms. The fraction of the number of images within a dataset is plotted with respect to AEE.

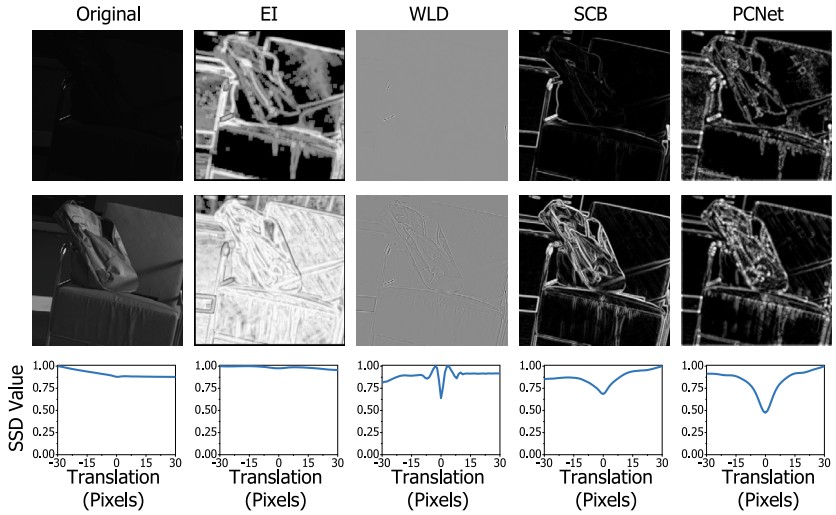


Fig. 11. Comparison of our PCNet and other similarity enhancement algorithms on scene *S1*. First row: the original and transformed reference images. Second row: the original and transformed floating images. Third row: the SSD distributions with respect to the horizontal translation from −30 to 30 pixels.

floating image and then displaying them in false color. It is observed that for both scenes, only PCNet produces registration results stably and accurately. It is worth noting that LGHD and RIFT generally produce roughly right registration results yet lack precision.

5.5. Comparison with deep-learning registration methods

In this subsection, we compare our PCNet with 4 state-of-the-art registration algorithms using deep CNNs, including DHN [32], MHN [34], UDHN [33]. The above networks are trained on the first 10 scenes in alphabetic order of the CAVE dataset and evaluated on the rest of the dataset, which is denoted as CAVE-CAVE. We then evaluate all the above methods on RGB/NIR dataset without retraining, which is denoted as CAVE-RGB/NIR, to compare their generalization ability. We note that our PCNet is only trained once on the first 10 scenes in alphabetic order of the CAVE dataset for a fair comparison.

We draw the fraction of the number of images with respect to ACE within CAVE and RGB/NIR datasets in Fig. 15. It is observed that in the CAVE-CAVE evaluation in Fig. 15(a) that all versions of PCNet outperform DHN and UDHN, but are surpassed by MHN when ACE is larger than around 2 pixels. The strong learning ability of deep neural networks contributes to the outstanding performance of MHN. We then focus on the cross dataset evaluation, namely the CAVE-RGB/NIR evaluation in Fig. 15(b). It is observed that PCNet outperforms all competitors, which indicates a relatively better generalization ability of the registration framework using PCNet.

An interesting phenomenon is observed in Fig. 15(a) that our proposed PCNet registration framework can achieve relatively higher accuracy than the deep-learning methods. On the contrary, in the case that ACE ranges from 2 pixels to 10 pixels, the deep-learning methods such as MHN performs better. The advantages of both methods can be combined by using our PCNet to boost the registration produced by the deep-learning methods as in [68].

We list the percentage of number of images under ACEs of 1, 5, and 10 pixels for each deep-learning method in Table 5. It is observed that our PCNet boosts the registration accuracy of the deep-learning methods via a considerably large gap. On the CAVE dataset, our PCNet raises the percentage under 1 pixel of UDHN from 0.1% to 82.5%. On the RGB/NIR dataset, PCNet boosts the percentage under 1 pixel of DHN from 0.0% to 63.5%.

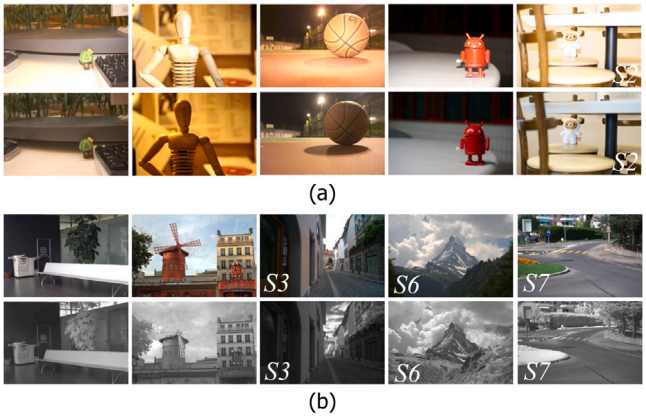


Fig. 12. Example images for RGB/NIR and flash/no-flash datasets. (a) RGB/NIR dataset (b) Flash/no-flash dataset. The scenes to be used for illustration are marked with *S2*, *S3*, *S6*, and *S7*.

Table 5
The percentage of number of images under different degrees of ACEs using deep-learning methods and their PCNet boosted versions.

ACE (pixels)	CAVE			RGB/NIR		
	< 1	< 5	< 10	< 1	< 5	< 10
DHN	0.0%	0.2%	4.3%	0.0%	0.0%	6.1%
DHN+PCNet	**67.0%**	**72.0%**	**72.3%**	**63.5%**	**74.2%**	**75.9%**
MHN	24.3%	74.0%	82.3%	9.6%	50.3%	65.5%
MHN+PCNet	**80.9%**	**86.9%**	**87.3%**	**68.4%**	**78.3%**	**81.5%**
UDHN	0.1%	24.6%	61.5%	0.0%	13.8%	41.4%
UDHN+PCNet	**82.5%**	**89.0%**	**89.1%**	**63.9%**	**74.0%**	**75.5%**

Table 6
Number of parameters and inference time of PCNet and other deep-learning registration algorithms.

	DHN [32]	MHN [34]	UDHN [33]	PCNet	PCNet-C2	PCNet-C1
# Parameters	34.19M	2.57M	21.29M	0.014M	0.0048M	0.0048M
Time (s)	0.002	0.012	0.015	1.298	0.482	0.270

Fig. 16 illustrates the registration results of scenes *S4*, *S5*, *S6*, and *S7* produced by deep-learning registration algorithms and their PCNet boosted versions. The white boxes highlight the details for comparison. It is observed that our PCNet can boost the varying degrees of registration results of deep-learning methods significantly. The obvious registration performance gap among the deep-learning methods (e.g. DHN and MHN) is removed after the boosting of our PCNet.

Efficiency Comparison. It is worth taking comparison of the efficiency of the deep-learning methods and PCNet. We first compare the network parameters and the inference time for DHN, UDHN, MHN, PCNet, PCNet-C2, and PCNet-C1 in Table 6. The comparison of inference time is conducted on the machine having an Intel(R) Core(TM) i9-11900 CPU @ 2.50 GHz with 16G memory and an NVIDIA Quadro RTX 8000. It is observed that all versions of PCNet have the advantage in network parameters, but lacks the superior in inference time. The number of parameters of PCNet is about 2400× less compared to DHN and 1500× to MHN, with PCNet-C2 being 7200× and 4500×. We note that the iterative optimization of PCNet is implemented in MATLAB code without any acceleration, run only on CPU, and calculated with the maximum iteration for each pyramid. The computational efficiency would be much improved if implemented using C++ with GPU acceleration.

We further compare the training efficiency of the above methods by reducing their original training iterations and image samples to 10% and 1%. We note that the original training iterations of PCNet are already fewer than the deep-learning methods. The registration results of CAVE and RGB/NIR datasets under the reduced training iterations and image samples are illustrated in Fig. 17. It is observed that PCNet can keep performing well, which means PCNet can be fast trained with very few training data. On the contrary, the performance of deep-learning based methods obviously degenerates as the reduction of training iterations and image samples.

Table 7
Error statistics produced by image registration using different settings of PCNet on the CAVE dataset. For brevity bestN% is denoted as B.N%. The best indicators are in bold.

Ablation part	Mean	Med.	Tri.	B.25%	B.50%	B.75%	B.95%
Freezing layer1	15.65	0.32	9.27	**0.07**	0.13	4.83	13.36
Freezing layer2	14.83	0.39	8.85	**0.07**	0.16	4.13	12.53
w/o LCK	14.90	0.31	9.06	0.08	0.14	3.93	12.61
w/o modification	14.93	**0.31**	8.79	**0.07**	**0.13**	4.22	12.66
w/o Gabor filter	21.93	20.44	20.79	0.38	2.28	12.49	19.93
Full	**13.33**	**0.31**	**7.67**	**0.07**	0.14	**2.72**	**10.98**

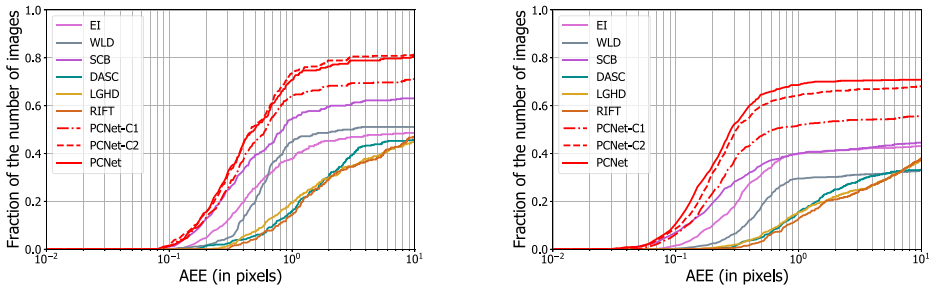
5.6. Comparison of the knowledge injection strategy

As mentioned, PCNet achieves similarity enhancement via the injection of phase congruency knowledge directly into the network architecture. We compare the knowledge injection strategy of PCNet to the similarity enhancement method in [68], namely deep Lucas–Kanade feature map (DLKFM), which injects knowledge to pure CNN network architecture by the designed loss function and feature map computing strategy. We compare the registration performance of PCNet and DLKFM on CAVE and RGB/NIR datasets in Fig. 18. We make the comparison fair by using PCNet-C1, which has 1 channel feature map as DLKFM. The same registration framework is adopted, also for a fair comparison. We train DLKFM on CAVE using the same training data as PCNet-C1. It is observed that PCNet-C1 owns a significantly better registration performance, which is likely to benefit from the direct knowledge injection into the network architecture.

5.7. Ablation study

We conduct the ablation study of our PCNet for the noise estimation layer (layer1), modified phase deviation estimation layer (layer2), learnable convolutional kernels (LCK), modification of Gabor filter, and Gabor filter. Table 7 lists the error statistics of our full PCNet, as well as the networks whose network parts are individually frozen or removed. It is observed that our full PCNet has the best registration performance. Specifically, the networks without the learnable convolutional kernel (LCK), the Gabor filter, or the modification of the Gabor filter are of evident performance degradation, which means the added network parts contribute a lot to the registration accuracy. The freezing of the modified phase deviation estimation layer (layer2) also causes a relatively large drawback in registration accuracy.

We further validate the necessity of phase congruency in the similarity enhancement task. We replace our PCNet with a U-net [69], and then train the network on the same loss Eq. (21) as PCNet. However, during training, the output feature maps of U-net using each of the above losses become meaningless patterns that lack information relating to the input images and cannot be used for image registration as illustrated in Fig. 19. It seems that without prior knowledge, it is unlikely to train the similarity enhancement network with the only similarity loss. The same conclusion has also been verified recently. In [68], the convolutional networks for similarity enhancement are trained by not only constraining the output feature maps to be similar using MSE loss, but also adding a convergence loss to guarantee the output feature maps have a smooth surface around the ground truth transformation.



(a) Evaluation on flash/no-flash dataset. (b) Evaluation on RGB/NIR dataset.

Fig. 13. Registration evaluation on flash/no-flash and RGB/NIR datasets using PCNet and other registration algorithms. The fraction of the number of images within a dataset is plotted with respect to AEE.

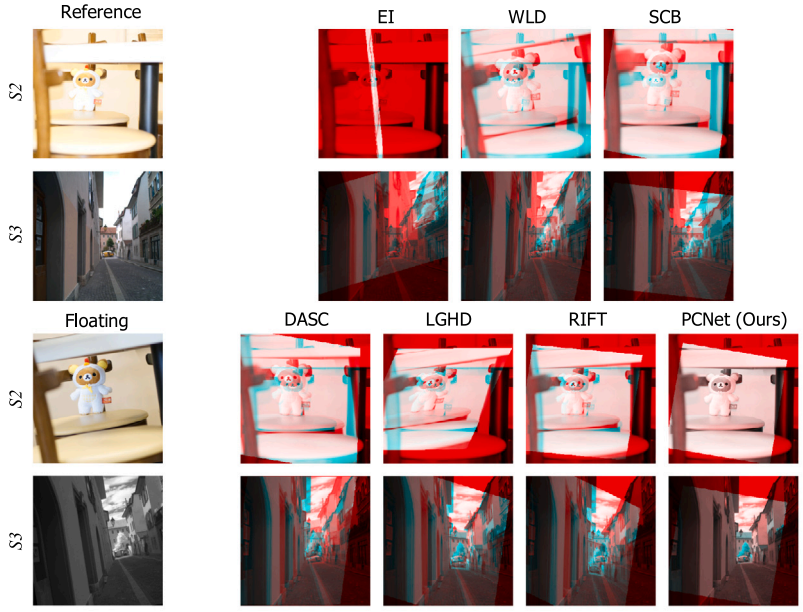
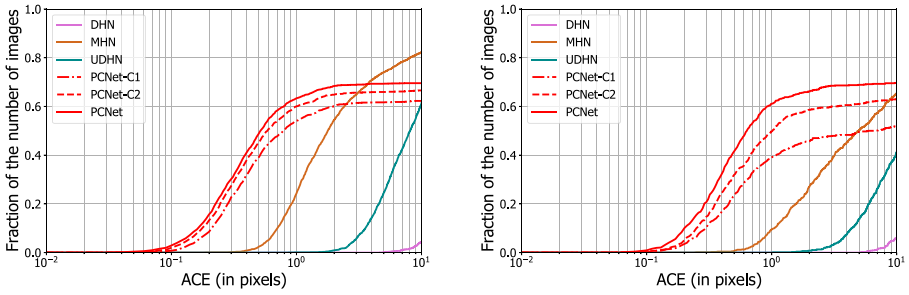


Fig. 14. Registration results produced by PCNet and other registration algorithms. First row: registration results of RGB/NIR image pair, scene *S2*. Second row: registration results of flash/no-flash image pairs, scene *S3*.



(a) CAVE-CAVE evaluation. (b) CAVE-RGB/NIR evaluation.

Fig. 15. Registration evaluation on CAVE and RGB/NIR datasets using PCNet and other deep-learning registration algorithms. The fraction of the number of images within a dataset is plotted with respect to ACE.

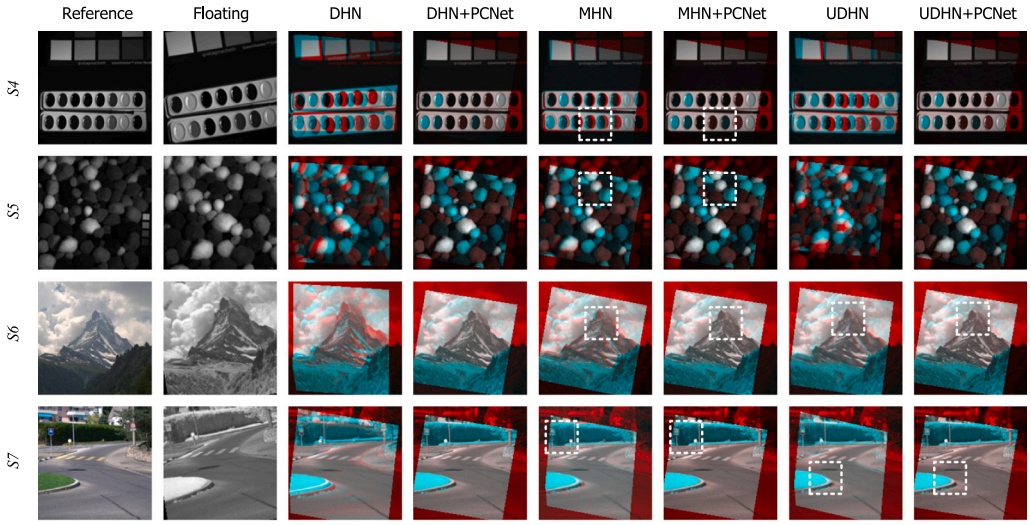


Fig. 16. Registration results produced by deep-learning registration algorithms together with their PCNet boosted version. First and second row: registration results of CAVE image pair, scene *S4* and *S5*. Third and fourth row: registration results of RGB/NIR image pairs, scene *S6* and *S7*. The white boxes highlight the details for comparison.

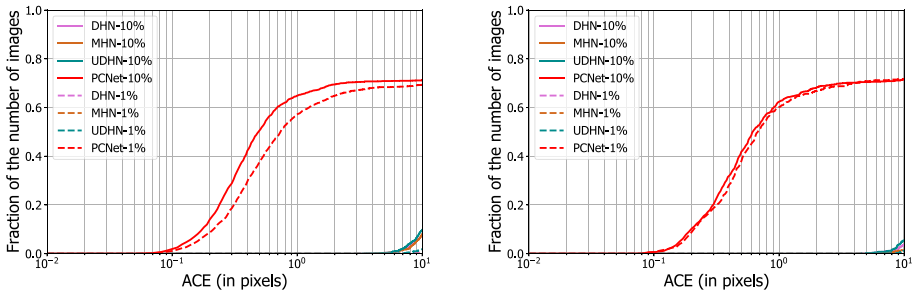


Fig. 17. Training efficiency evaluation on CAVE and RGB/NIR datasets of PCNet and other deep-learning registration algorithms. The fraction of the number of images within a dataset is plotted with respect to ACE.

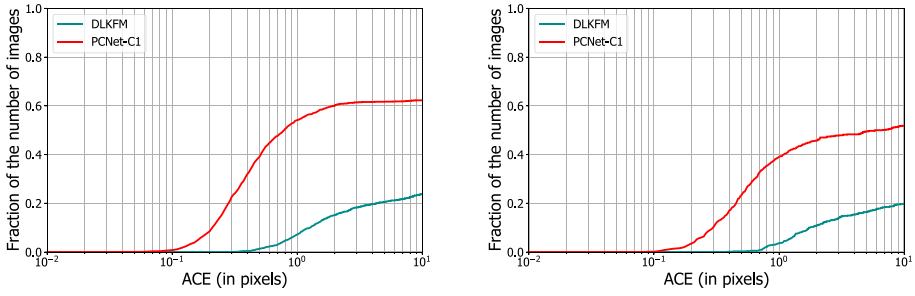


Fig. 18. Knowledge injection strategy comparison on CAVE and RGB/NIR datasets of the similarity enhancement methods including PCNet and DLKFM. The fraction of the number of images within a dataset is plotted with respect to ACE.

6. Conclusions

In this paper, we have proposed a network called PCNet to enhance the structure similarity for the purpose of image registration. The prior knowledge of our PCNet is based on phase congruency. PCNet is concise and easy to train thanks to the prior information. It produces satisfactory registration results on a variety of multispectral and multimodal datasets though only trained on a subset of the CAVE multispectral dataset. The PCNet performs better than other state-of-the-art similarity enhancement algorithms and feature-based registration algorithms.

We note that currently our PCNet is directly combined with a traditional registration framework. This direct combination makes the trainable part of the network lightweight but may restrain the registration performance. It would be our future work to transform the current PCNet into a flexible structure that can be easily incorporated into trainable deep-learning registration networks. Considering that in the real world multispectral and multimodal images are usually unregistered, it is also worthwhile to explore the possibility of training PCNet on unregistered images.

Fig. 19. Similarity enhanced feature maps generated by U-net. (a) structure protection parameter $c = 0.7$. (b) structure protection parameter $c = 2$.

CRediT authorship contribution statement

Si-Yuan Cao: Conceptualization, Methodology, Writing – original draft, Software. **Beinan Yu:** Methodology, Software, Writing – review & editing. **Lun Luo:** Investigation, Validation, Writing – review & editing. **Runmin Zhang:** Software, Validation, Writing – review & editing. **Shu-Jie Chen:** Project administration. **Chunguang Li:** Supervision, Writing – review & editing. **Hui-Liang Shen:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

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

The data used in the paper is publicly available. The code will be released once the manuscript is accepted.

Acknowledgment

This work was supported by the “Pioneer” and “Leading Goose” R & D Program of Zhejiang under grant 2023C03136.

References

[1] M. Brown, S. Süsstrunk, Multi-spectral SIFT for scene category recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2011, pp. 177–184.
[2] S. He, R.W. Lau, Saliency detection with flash and no-flash image pairs, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2014, pp. 110–124.
[3] F. Yasuma, T. Mitsunaga, D. Iso, S.K. Nayar, Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum, IEEE Trans. Image Process. 19 (9) (2010) 2241–2253.
[4] A. Chakrabarti, T. Zickler, Statistics of real-world hyperspectral images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2011, pp. 193–200.
[5] D. Guan, Y. Cao, J. Yang, Y. Cao, M.Y. Yang, Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection, Inf. Fusion 50 (2019) 148–157.
[6] Y. Cao, X. Luo, J. Yang, Y. Cao, M.Y. Yang, Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection, Inf. Fusion 88 (2022) 1–11.
[7] F.-P. An, J.-e. Liu, Pedestrian re-identification algorithm based on visual attention-positive sample generation network deep learning model, Inf. Fusion 86 (2022) 136–145.
[8] M. Müller, W. Krüger, G. Saur, Robust image registration for fusion, Inf. Fusion 8 (4) (2007) 347–353.
[9] Y. Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward, X. Wang, Deep learning for pixel-level image fusion: Recent advances and future prospects, Inf. Fusion 42 (2018) 158–173.
[10] G.M. Dimitri, S. Spasov, A. Duggento, L. Passamonti, P. Lió, N. Toschi, Multimodal and multicontrast image fusion via deep generative models, Inf. Fusion 88 (2022) 146–160.
[11] S. Xu, J. Zhang, J. Wang, K. Sun, C. Zhang, J. Liu, J. Hu, A model-driven network for guided image denoising, Inf. Fusion 85 (2022) 60–71.
[12] X. Guo, Y. Yang, C. Wang, J. Ma, Image dehazing via enhancement, restoration, and fusion: A survey, Inf. Fusion 86 (2022) 146–170.
[13] X. Shen, L. Xu, Q. Zhang, J. Jia, Multi-modal and multi-spectral registration for natural images, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2014, pp. 309–324.
[14] S.-J. Chen, H.-L. Shen, C. Li, J.H. Xin, Normalized total gradient: A new measure for multispectral image registration, IEEE Trans. Image Process. 27 (3) (2017) 1297–1310.
[15] B. Zitova, J. Flusser, Image registration methods: A survey, Image Vis. Comput. 21 (11) (2003) 977–1000.
[16] 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.
[17] V.A. Zimmer, M.Á.G. Ballester, G. Piella, Multimodal image registration using Laplacian commutators, Inf. Fusion 49 (2019) 130–145.
[18] S.-Y. Cao, H.-L. Shen, S.-J. Chen, C. Li, Boosting structure consistency for multispectral and multimodal image registration, IEEE Trans. Image Process. 29 (2020) 5147–5162.
[19] 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 (2021) 1–5.
[20] C. Wachinger, N. Navab, Structural image representation for image registration, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition-Workshops, CVPRW, IEEE, 2010, pp. 23–30.
[21] M.C. Morrone, R.A. Owens, Feature detection from local energy, Pattern Recognit. Lett. 6 (5) (1987) 303–313.
[22] P. Kovesi, Phase congruency: A low-level image invariant, Psychol. Res. 64 (2) (2000) 136–148.
[23] Z. Liu, R. Laganière, Phase congruence measurement for image similarity assessment, Pattern Recognit. Lett. 28 (1) (2007) 166–172.
[24] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Trans. Image Process. 20 (8) (2011) 2378–2386.
[25] K. Wang, P. Xiao, X. Feng, G. Wu, Image feature detection from phase congruency based on two-dimensional Hilbert transform, Pattern Recognit. Lett. 32 (15) (2011) 2015–2024.
[26] X. Yuan, P. Shi, Iris feature extraction using 2D phase congruency, in: Third International Conference on Information Technology and Applications, Vol. 2, ICITA’05, IEEE, 2005, pp. 437–441.
[27] G. Bhatnagar, Q.J. Wu, Z. Liu, Directive contrast based multimodal medical image fusion in NSCT domain, IEEE Trans. Multimed. 15 (5) (2013) 1014–1024.
[28] H. Li, H. Qiu, Z. Yu, Y. Zhang, Infrared and visible image fusion scheme based on NSCT and low-level visual features, Infrared Phys. Technol. 76 (2016) 174–184.
[29] A. Wong, D.A. Clausi, ARRSI: Automatic registration of remote-sensing images, IEEE Trans. Geosci. Remote Sens. 45 (5) (2007) 1483–1493.
[30] Y. Ye, J. Shan, L. Bruzzone, L. Shen, Robust registration of multimodal remote sensing images based on structural similarity, IEEE Trans. Geosci. Remote Sens. 55 (5) (2017) 2941–2958.
[31] J. Li, Q. Hu, M. Ai, RIFT: Multi-modal image matching based on radiation-variation insensitive feature transform, IEEE Trans. Image Process. 29 (2019) 3296–3310.
[32] D. DeTone, T. Malisiewicz, A. Rabinovich, Deep image homography estimation, 2016, arXiv preprint arXiv:1606.03798.
[33] J. Zhang, C. Wang, S. Liu, L. Jia, N. Ye, J. Wang, J. Zhou, J. Sun, Content-aware unsupervised deep homography estimation, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2020, pp. 653–669.
[34] H. Le, F. Liu, S. Zhang, A. Agarwala, Deep homography estimation for dynamic scenes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 7652–7661.
[35] J. Ma, X. Jiang, A. Fan, J. Jiang, J. Yan, Image matching from handcrafted to deep features: A survey, Int. J. Comput. Vis. (2020) 1–57.
[36] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, Multimodality image registration by maximization of mutual information, IEEE Trans. Med. Imaging 16 (2) (1997) 187–198.
[37] Y. Hel-Or, H. Hel-Or, E. David, Matching by tone mapping: Photometric invariant template matching, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 317–330.
[38] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, W. Gao, WLD: A robust local image descriptor, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2009) 1705–1720.
[39] G.T. Fechner, D.H. Howes, E.G. Boring, Elements of Psychophysics, Vol. 1, Holt, Rinehart and Winston New York, 1966.
[40] F. Yang, M. Ding, X. Zhang, Y. Wu, J. Hu, Two phase non-rigid multi-modal image registration using weber local descriptor-based similarity metrics and normalized mutual information, Sensors 13 (6) (2013) 7599–7617.
[41] 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.
[42] C.A. Aguilera, A.D. Sappa, R. Toledo, LGHD: A feature descriptor for matching across non-linear intensity variations, in: Proceedings of the IEEE International Conference on Image Processing, ICIP, 2015, pp. 178–181.
[43] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[44] Y. Liao, Y. Di, H. Zhou, A. Li, J. Liu, M. Lu, Q. Duan, Feature matching and position matching between optical and SAR with local deep feature descriptor, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15 (2021) 448–462.
[45] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[46] 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.
[47] Y. Ye, L. Bruzzone, J. Shan, F. Bovolo, Q. Zhu, Fast and robust matching for multimodal remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 57 (11) (2019) 9059–9070.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.
[49] M.C. Morrone, A. Navangione, D. Burr, An adaptive approach to scale selection for line and edge detection, Pattern Recognit. Lett. 16 (7) (1995) 667–677.
[50] B. Robbins, R. Owens, 2D feature detection via local energy, Image Vis. Comput. 15 (5) (1997) 353–368.
[51] J.-C. Yoo, T.H. Han, Fast normalized cross-correlation, Circuits Systems Signal Process. 28 (6) (2009) 819–843.
[52] H. Liu, Z. Fu, J. Han, L. Shao, S. Hou, Y. Chu, Single image super-resolution using multi-scale deep encoder–decoder with phase congruency edge map guidance, Inform. Sci. 473 (2019) 44–58.
[53] J. Pan, J. Dong, J.S. Ren, L. Lin, J. Tang, M.-H. Yang, Spatially variant linear representation models for joint filtering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1702–1711.
[54] D. Varga, No-reference image quality assessment based on the fusion of statistical and perceptual features, J. Imaging 6 (8) (2020) 75.
[55] M. Wang, C. Sun, A. Sowmya, Complex shearlets and rotary phase congruence tensor for corner detection, Pattern Recognit. 128 (2022) 108606.
[56] B. Cyganek, Object Detection and Recognition in Digital Images: Theory and Practice, John Wiley & Sons, 2013.
[57] R. Reams, Hadamard inverses, square roots and products of almost semidefinite matrices, Linear Algebra Appl. 288 (1999) 35–43.
[58] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the International Conference on International Conference on Machine Learning, ICML, Omni Press, 2010, pp. 807–814.
[59] D. Gabor, Theory of communication. Part 1: The analysis of information, J. Inst. Electr. Eng.-Part III: Radio Commun. Eng. 93 (26) (1946) 429–441.
[60] S. Marčelja, Mathematical description of the responses of simple cortical cells, JOSA 70 (11) (1980) 1297–1300.
[61] S. Luan, C. Chen, B. Zhang, J. Han, J. Liu, Gabor convolutional networks, IEEE Trans. Image Process. 27 (9) (2018) 4357–4366.
[62] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[63] S. Abbasi, F. Mokhtarian, Shape similarity retrieval under affine transform: Application to multi-view object representation and recognition, in: Proceedings of the IEEE International Conference on Computer Vision, Vol. 1, ICCV, 1999, pp. 450–455.
[64] J. Klein, T. Aach, Multispectral filter wheel cameras: Modeling aberrations for filters in front of lens, in: Digital Photography VIII, Vol. 8299, International Society for Optics and Photonics, 2012, p. 82990R.
[65] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for image restoration with neural networks, IEEE Trans. Comput. Imaging 3 (1) (2016) 47–57.
[66] C. Liu, J. Yuen, A. Torralba, J. Sivic, W.T. Freeman, SIFT flow: Dense correspondence across different scenes, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, 2008, pp. 28–42.
[67] P.H. Torr, A. Zisserman, MLESAC: A new robust estimator with application to estimating image geometry, Comput. Vis. Image Underst. 78 (1) (2000) 138–156.
[68] Y. Zhao, X. Huang, Z. Zhang, Deep Lucas-Kanade homography for multimodal image alignment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 15950–15959.
[69] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, Springer, 2015, pp. 234–241.