Dense Cross-Modal Correspondence Estimation with the Deep Self-Correlation Descriptor

Seungryong Kim, *Member, IEEE,* Dongbo Min, *Senior Member, IEEE,* Stephen Lin, *Member, IEEE,* and Kwanghoon Sohn, *Senior Member, IEEE*

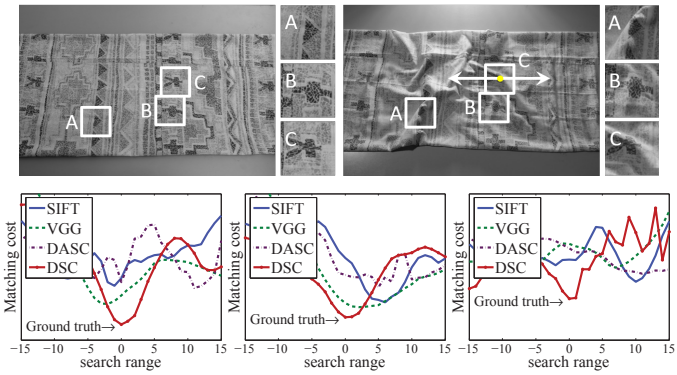
Abstract—We present the deep self-correlation (DSC) descriptor for establishing dense correspondences between images taken under different imaging modalities, such as different spectral ranges or lighting conditions. We encode local self-similar structure in a pyramidal manner that yields both more precise localization ability and greater robustness to non-rigid image deformations. Specifically, DSC first computes multiple self-correlation surfaces with randomly sampled patches over a local support window, and then builds pyramidal self-correlation surfaces through average pooling on the surfaces. The feature responses on the self-correlation surfaces are then encoded through spatial pyramid pooling in a log-polar configuration. To better handle geometric variations such as scale and rotation, we additionally propose the geometry-invariant DSC (GI-DSC) that leverages multi-scale self-correlation computation and canonical orientation estimation. In contrast to descriptors based on deep convolutional neural networks (CNNs), DSC and GI-DSC are training-free (i.e., handcrafted descriptors), are robust to cross-modality, and generalize well to various modality variations. Extensive experiments demonstrate the state-of-the-art performance of DSC and GI-DSC on challenging cases of cross-modal image pairs having photometric and/or geometric variations.

Index Terms—Cross-modal correspondence, pyramidal structure, self-correlation, local self-similarity, non-rigid deformation

1 Introduction

In many computer vision and computational photography applications, images captured under different imaging modalities supplement the data provided in color images. Typical examples of other imaging modalities include infrared [1], [2], [3] and dark flash [4] photography. More broadly, photos taken under different imaging conditions, such as exposure settings [5], blur levels [6], [7], and illumination [8], can also be considered as cross-modal [9], [10].

Establishing dense correspondences between such cross-modal image pairs is essential for combining their disparate information. However, basic visual properties, including color or gradients, are frequently not shared across cross-modal images, thus degrading matching by conventional feature descriptors [11], [12]. Moreover, geometric variations frequently appear among them taken under different viewpoints or containing moving objects. Although powerful global optimizers can help to improve the accuracy of correspondence estimation to some extent [13], [14], inherent limitations exist without suitable matching descriptors [15]. For instance, scale invariant feature transform (SIFT) [11], one of the most popular feature descriptors, provides relatively good matching performance when there are small photometric and geometric variations, but it frequently fails to capture reliable matching evidence across cross-modal images due to their different visual properties [9], [10].



(a) Cost profile of A (b) Cost profile of B (c) Cost profile of C

Fig. 1. Examples of matching cost profiles, computed with descriptors such as SIFT [11], VGG-Net (conv3-4) [16], DASC [10], and DSC along the scan lines of A, B, and C for image pairs under non-rigid deformations and illumination changes. In comparison to other handcrafted and deep CNN-based descriptors, DSC yields more reliable global minima.

Although convolutional neural network (CNN) based features [17], [18], [19], [20], [21], [22], [23], [24] have recently emerged as a robust alternative, they cannot satisfactorily address severe cross-modal variations, since their shared and fixed convolutional kernels across cross-modal images often produce inconsistent feature maps [21], [25]. Of particular importance, there lacks a cross-modal benchmark with dense ground-truth correspondences, making supervised learning of CNNs less feasible for this task. In addition, a network trained on small-scale datasets may be overfitted

- *S. Kim is with the School of Computer and Communication Sciences (IC), École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland.*
  E-mail: seungryong.kim@epfl.ch
- *D. Min is with the Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, South Korea.*
  E-mail: dbmin@ewha.ac.kr
- *S. Lin is with Microsoft Research Asia, Beijing 100080, China.*
  E-mail: stevelin@microsoft.com
- *K. Sohn is with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea.*
  E-mail: khsohn@yonsei.ac.kr

** Corresponding author*

to specific modalities.

Meanwhile, to address the problem of cross-modal appearance and shape changes, feature descriptors have been proposed to leverage local self-similarity (LSS) [26], which is motivated by the notion that the geometric layout of local self-similarities is relatively insensitive to visual property variations. The descriptor, called dense adaptive self-correlation (DASC) [10], makes use of LSS and has demonstrated high accuracy on cross-modal image pairs. However, DASC suffers from two significant shortcomings. One is its limited discriminative power due to a limited set of patch sampling patterns used for modeling internal self-similarities. The other major shortcoming is that DASC does not provide the flexibility to deal with non-rigid geometric deformations, which deteriorates the matching accuracy. More recently, a fully convolutional self-similarity (FCSS) descriptor [24] was proposed to formulate LSS within a deep network. However, its application to cross-modal correspondence has not been studied.

In this paper, we present a descriptor, called deep self-correlation (DSC), that overcomes the shortcomings of previous LSS-based descriptors [10], [24], [26] and provides robust cross-modal correspondence. This work is motivated by the observation that local self-similarity appears in a multi-scale fashion, and thus it is formulated with a pyramidal structure that enhances localization ability and robustness to photometric and geometric deformations. Unlike LSS [26] which computes self-similarity with respect to only a central patch, and DASC [10] which selects different patch pairs and calculates the self-similarity between them, DSC computes self-correlation surfaces representing the self-similarity between randomly selected patches and all other patches, and then aggregates these responses to more comprehensively encode structural information. This aggregation of self-similarity responses is performed through log-polar spatial pyramid pooling (L-SPP) where a support window is partitioned into log-polar divisions from coarse to fine levels, which yields a representation less sensitive to non-rigid deformations. For efficient computation of DSC over densely sampled pixels, we calculate the self-correlation surfaces through fast edge-aware filtering.

Furthermore, to better address geometric variations, we propose the geometry-invariant DSC descriptor, called GI-DSC. In formulating this extension, we leverage the assumption that geometric deformation fields can be approximated locally by a similarity transformation (i.e., translation, rotation, and uniform scaling). Specifically, to deal with scale deformations, multi-scale self-correlation surfaces are first measured on the image pyramid, and then used to encode maximal self-similarities across scales, which remain consistent to scale changes. Canonical orientations are also estimated with the maximum orientation bin weighted by the self-correlation values.

Compared to existing CNN-based descriptors [17], [18], [19], [20], [21], [22], [23] as well as FCSS [24], DSC requires no training data and thus generalizes well to various modality variations. Fig. 1 illustrates the robustness of DSC for image pairs with non-rigid geometric deformations and illumination changes in comparison to existing handcrafted and even deep CNN-based methods [10], [11], [16].

In experimental results, we show that DSC outperforms existing feature descriptors and similarity measures on various benchmarks containing photometric and/or geometric variations: (1) the Middlebury stereo benchmark [27] with illumination and exposure variations; (2) a cross-modal and cross-spectral dataset [9], [10] including RGB and near-infrared (NIR) images [1], [9], different exposures [5], [9], flash-noflash images [8], blurry images [6], [7], and RGB-depth images [9]; (3) the DaLI benchmark [28] containing non-rigid geometric deformations; (4) the tri-modal human body segmentation benchmark [29] including RGB, depth, and far-infrared (FIR) images; and (5) the DIML benchmark [30] including RGB images with both photometric and geometric variations.

This manuscript extends the conference version [31] through (1) a geometry-invariant extension of DSC, called GI-DSC; (2) an in-depth analysis of DSC and GI-DSC; and (3) an extensive comparative study with state-of-the-art CNN-based descriptors using various datasets. The source code will be available online at our project webpage: `https://seungryong.github.io/DSC/`.

2 Related Work

2.1 Handcrafted and Learned Feature Descriptors

Conventional gradient-based descriptors such as SIFT [11] or DAISY [12], as well as intensity comparison-based binary descriptors such as BRIEF [32], have shown limited performance for estimating dense correspondences between cross-modal image pairs. Several attempts have been made using machine learning algorithms to derive features from large-scale datasets [17], [33]. Recently, for designing feature descriptors based on a CNN architecture, intermediate activations are extracted as the descriptor [17], [18], [19], [20], [21], [22], [23], [24], showing effectiveness for local matching. However, even though CNN-based descriptors encode a discriminative structure, they have inherent limitations for cross-modal image correspondence because they are derived from convolutional layers using shared kernels [21], [25]. Furthermore, the dearth of ground-truth data for cross-modal correspondence presents an obstacle for supervised learning of CNNs in this context.

For cross-modal correspondence, variants of SIFT have been developed [34], but like SIFT they maintain an inherent limitation in dealing with gradients that vary differently between modalities. For illumination invariant correspondence, Wang et al. proposed the local intensity order pattern (LIOP) descriptor [35], but radiometric variations often alter the relative order of pixel intensities. Simo-Serra et al. proposed the deformation and light invariant (DaLI) descriptor [28] to provide high resilience to non-rigid transformations and illumination changes, but it in practice cannot provide dense descriptors in the image domain due to its heavy computational load. Recently, CNN-based cross-spectral similarity models [36], [37] have shown improved performance on RGB-NIR correspondence, but they require supervised learning, thus limiting its applicability to various cross-modal correspondence tasks.

Schechtman and Irani introduced the local self-similarity (LSS) descriptor [26] for the purpose of template matching, and achieved impressive results. By employing LSS, many

approaches have tried to solve for cross-modal correspondence [38], [39], [40]. However, none of these approaches scale well to dense correspondence due to limited discriminative power and high complexity. Inspired by LSS, Kim et al. proposed DASC [10] to estimate cross-modal dense correspondences, but it is not able to handle non-rigid deformations and has limited discriminative power due to its fixed patch pooling scheme. More recently, FCSS [24] formulated LSS within a fully convolutional network where patch sampling patterns and self-similarity measure are both learned. Although FCSS improved performance dramatically for semantic correspondence, it is tailored to object-level correspondence, instead of cross-modal image pairs at a scene level. Moreover, it cannot deal with severe geometric variations which frequently appear across cross-modal images.

2.2 Area-Based Similarity Measures

A popular method for medical image registration is mutual information (MI) [41], but the variations it can reliably handle are only of global transformations. [42] alleviates this issue by leveraging a locally adaptive weight obtained from SIFT matching, but its performance is still limited on cross-modal variation [43]. Although cross-correlation based methods such as adaptive normalized cross-correlation (ANCC) [44] produce satisfactory results for locally linear variations, they are less effective against more substantial modality variations. Irani et al. employed cross-correlation on a Laplacian energy map for measuring multi-sensor image similarity [45], but this exhibits limited performance in general image matching tasks. Shen et al. proposed robust selective normalized cross-correlation (RSNCC) [9] for dense alignment between cross-modal images, but as an intensity based measure it can still be sensitive to cross-modal variations. DeepMatching [46] was proposed to compute dense correspondences by employing a hierarchical pooling scheme like in a CNN, but it is not designed to handle cross-modal matching.

2.3 Geometry-Invariant Correspondence Estimation

To alleviate geometric variation problems, many methods have been proposed based on SIFT flow (SF) [13] optimization, including deformable spatial pyramid (DSP) [14], scale-less SIFT flow (SLS) [47], scale-space SIFT flow (SSF) [48], and generalized DSP (GDSP) [49]. However, the large search spaces for establishing geometry-invariant dense correspondence make computational complexity a critical limitation of these methods. Barnes et al. proposed generalized PatchMatch (GPM) [50] for efficient matching based on a randomized search scheme. Yang et al. proposed DAISY Filter Flow (DFF) [51], which utilizes the DAISY descriptor [12] with the PatchMatch Filter (PMF) [52], to provide geometric invariance. However, its weak spatial smoothness often induces mismatched results. While the aforementioned methods have attempted to address the problem from an optimization perspective, various geometry-invariant descriptors have also been developed for geometry-invariant correspondence estimation. Kokkinos et al. proposed the scale invariant descriptor (SID) [53] to encode geometric robustness in the descriptor itself, but it does not deal with multi-modal matching. A segmentation-aware approach [54] was presented to provide geometric robustness for descriptors, e.g., SIFT [11] or SID [53], but it can have a negative effect on the discriminative power of the descriptor. More recently, geometry-invariant DASC (GI-DASC) [30] employed DASC in a superpixel-based representation with estimated geometric fields. Although it provides improved robustness to geometric variations, it inherits the limitations of DASC, and its performance is sensitive to superpixel segmentation.

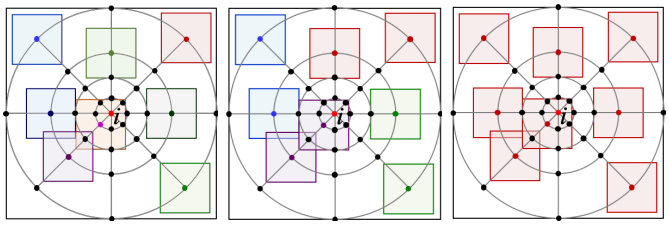


Fig. 2. Illustration of DSC that uses log-polar spatial pyramid pooling on pyramidal self-similarity surfaces defined at (a) level 3, (b) level 2, and (c) level 1. Different colors represent different patches used to reconstruct the self-correlation surfaces while same colors represent patches in the same set used in an aggregation procedure.

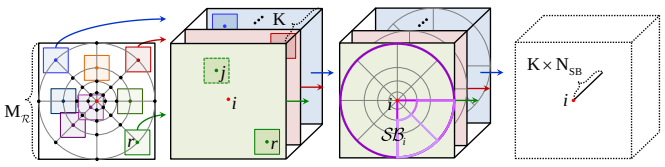
3 Background and Overview

Unlike conventional descriptors [11], [12] that rely on basic visual properties such as color or gradients, LSS-based descriptors represent local self-similar structures by recording the similarity between certain patch pairs based on the observation that the geometric layout of the local self-similarities is preserved across cross-modal image pairs [10], [24], [26]. Formally, given an image f_i for pixel i, LSS descriptor $\mathcal{D}_i = \{d_i(l)\}$ is defined on a local support window $\mathcal{R}_i$ for $l \in \{1, ..., L\}$ with the feature dimension L such that

$$d_i(l) = \max_{t \in \mathcal{T}_i(l)} \exp(-\mathcal{S}(s_i(l), t)/\sigma_c). \quad (1)$$

$\mathcal{S}(s, t)$ is a self-similarity distance between two local patches sampled on pixels s and t. $s_i(l)$ and $\mathcal{T}_i(l)$ are l-th anchor point and pooling bin, respectively. To alleviate the effects of outliers, the self-similarity responses are encoded by non-linear mapping with an exponential function of bandwidth σ_c. For spatial invariance to the position of the sampling pattern, the maximum self-similarity within the pooling bin $\mathcal{T}_i(l)$ is computed. Based on this general framework, LSS has been formulated in various ways, using different self-similarity distances and different sampling strategies for the patch pairs [10], [24], [26].

As shown in Fig. 2(a), LSS [26] first computes a self-similarity surface, discretizes the surface into log-polar bins, and then stores the maximum value of each bin. It formally defines $s_i(l)$ as a fixed center pixel i and $\mathcal{T}_i(l)$ as a log-polar bin $\mathcal{B}_i(l)$, defined such that $\{j | j \in \mathcal{R}_i, \rho_{r-1} < |i-j| \leq \rho_r, \phi_{a-1} < \angle(i-j) \leq \phi_a\}$ with a log radius ρ_r for $r \in \{1, \cdots, N_\rho\}$ and a quantized angle ϕ_a for $a \in \{1, \cdots, N_\phi\}$ with $\rho_0 = 0$ and $\phi_0 = 0$, where each pair of r and a is associated with a unique index l. $\mathcal{S}(\cdot, \cdot)$ is computed using the sum of squared differences (SSD) [26]. Though LSS provides robustness to modality variations, matching details are not well preserved and its significant computation does not scale well for estimating dense correspondences.



(a) (b) (c) (d)

Fig. 3. Computation of single self-correlation (SSC) descriptor for (a) a local support window with random samples. (b) For each random patch, it first computes the self-similarity using an adaptive self-correlation measure, building multiple self-correlation surfaces. (c) It then encodes responses on the surfaces through log-polar spatial pyramid pooling. (d) The responses are concatenated into a feature vector.

DASC [10] encodes a set of the self-similarities between patch pairs randomly sampled from a log-polar point $\mathcal{P}_i(l)$ as shown in Fig. 2(b), defined such that $\{j|j \in \mathcal{R}_i, |i-j| = \rho_r, \angle(i-j) = \phi_a\}$, which has a higher density of points near the center pixel, similar to DAISY [12]. DASC formally defines $s_i(l)$ as the l-th randomly sampled pixel and $\mathcal{T}_i(l)$ as the l-th paired sample pixel, which is a special case of the pooling bin with the size 1×1. $\mathcal{S}(\cdot,\cdot)$ is computed using an adaptive self-correlation measure inspired by [44]. Although the DASC descriptor provides satisfactory results for dense cross-modal correspondence estimation, its randomized receptive field pooling has limited representation power and does not accommodate non-rigid deformations.

Inspired by DASC [10], our DSC descriptor also utilizes an adaptive self-correlation measure between two patches. However, we adopt a different strategy in a manner that builds pyramidal self-similarity surfaces through the aggregation of multiple self-correlation responses on a single level to improve localization ability and robustness to non-rigid deformation. First of all, to compute multiple self-similarities, we formally define the anchor point $s_i(l)$ as the l-th randomly sampled pixel and the pooling bin $\mathcal{T}_i(l)$ as a log-polar pyramidal bin (Sec. 4.1). Moreover, to form pyramidal self-similarity surfaces, we also utilize average pooling, where the anchor point $s_i(l)$ is set to multiple points within a log-polar pyramidal point (Sec. 4.3). Finally, we alleviate problems caused by geometric variations, i.e., scale and/or rotation, in building the GI-DSC descriptor (Sec. 4.4). Fig. 2(c) illustrates the DSC descriptor, which incorporates log-polar spatial pyramid pooling on pyramidal self-correlation surfaces.

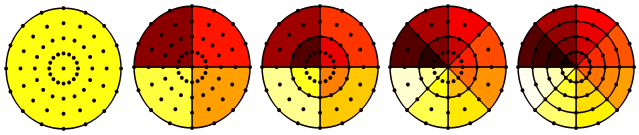
4 The DSC Descriptor

4.1 SSC: Single Self-Correlation

To overcome the limitations of self-similarity in the LSS [26] and DASC [10] descriptors, our approach builds pyramidal self-similarity surfaces, where feature responses are obtained through log-polar spatial pyramid pooling. We start by describing a single-scale version of DSC, which we refer to as single self-correlation (SSC).

4.1.1 Multiple Self-Correlations

Computing the local self-similarity with a single patch as in [26] is vulnerable to imaging deformations. To overcome this, we build multiple self-correlation surfaces. Specifically, we randomly select K points from a log-polar point set $\mathcal{P}_i$ defined within a local support window as in Fig. 3(a). We



(a) $s=1$ (b) $s=2$ (c) $s=3$ (b) $s=4$ (e) $s=5$

Fig. 4. Examples of log-polar pyramidal bins $\mathcal{SB}$. The total number of bins is $N_{\mathcal{SB}} = \sum_{s=2}^{S} 2^s + 1$, where S represents the pyramid level.

then convolve a patch $\mathcal{F}_r$ centered at the pixel r with all of patches $\mathcal{F}_j$ for $j \in \mathcal{R}_i$ as in Fig. 3(b). Similar to DASC [10], the similarity $\mathcal{C}(r,j)$ between patch pairs is measured using an adaptive self-correlation, which is known to be effective in addressing cross-modal variations, as follows:

$$\mathcal{C}(r,j) = \frac{\sum_{r',j'} \omega_{r,r'}\omega_{j,j'}(f_{r'} - G_r)(f_{j'} - G_j)}{\sqrt{\sum_{r'} \{\omega_{r,r'}(f_{r'} - G_r)\}^2} \sqrt{\sum_{j'} \{\omega_{j,j'}(f_{j'} - G_j)\}^2}}, \tag{2}$$

where $G_r = \sum_{r'} \omega_{r,r'} f_{r'}$ and $G_j = \sum_{j'} \omega_{j,j'} f_{j'}$ represent weighted intensity averages in pixels $r' \in \mathcal{F}_r$ and $j' \in \mathcal{F}_j$, respectively. Similar to DASC [10], the weight $\omega_{r,r'}$ represents how similar two pixels r and r' are, and the weight is normalized, i.e., $\sum_{r'} \omega_{r,r'} = 1$. It may be defined using any form of edge-aware weighting [55], [56], which increases the precision in describing self-similarities and boosts performance.

4.1.2 Log-polar Spatial Pyramid Pooling (L-SPP)

To encode the feature responses on the self-correlation surface, we exploit spatial pyramid pooling (SPP) [25], [57], [58], [59], [60], which has been shown to be robust to geometric deformations. We formulate this in a log-polar configuration, called log-polar spatial pyramid pooling (L-SPP). Note that some other descriptors also adopt log-polar pooling, which brings greater robustness because of its higher pixel density near the central pixel [12], [26], [32]. We also encode more structure information with LP-SPP.

Specifically, as shown in Fig. 4, the log-polar pyramidal bins $\mathcal{SB}_i(u)$ are first defined from the log-polar bins $\mathcal{B}_i(l)$, where u indexes all bins in all pyramidal levels $s \in \{1,...,S\}$ with the number of levels S. The log-polar pyramidal bin at the top of the pyramid, i.e., $s=1$, encompasses all of the bins $\mathcal{B}_i(l)$. The second level, i.e., $s=2$, is defined by dividing the top one into quadrants. For lower pyramid levels, i.e., $s > 2$, they are defined differently according to whether s is odd or even. For an odd s, the bins are defined by dividing bins in the upper level into two parts along the radius. For an even s, they are defined by dividing bins in the upper level into two parts with respect to angle. Thus, the number of log-polar pyramid bins is defined as $N_{\mathcal{SB}} = \sum_{s=2}^{S} 2^s + 1$.

As illustrated in Fig. 3(c), the feature responses are finally max-pooled on the log-polar pyramidal bins $\mathcal{SB}_i(u)$ of each self-correlation surface $\mathcal{C}(r,j)$, yielding the following feature response:

$$g_i(r,u) = \max_{j \in \mathcal{SB}_i(u)} \mathcal{C}(r,j). \tag{3}$$

It is repeated for all $r \in \{1,...,K\}$ and $u \in \{1,...,N_{\mathcal{SB}}\}$, yielding accumulated correlation responses $g_i^{\text{ssc}}(l) = \bigcup_{\{r,u\}} g_i(r,u)$ where l indexes over all r and u.

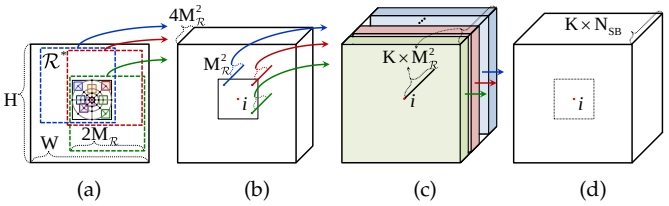


Fig. 5. Efficient computation of multiple self-similarity surfaces in an image: (a) An image with a doubled support window and random samples. (b) A 1-D vector representation of a self-similarity surface. (c) Self-similarity surfaces. (d) Self-similarity responses after L-SPP. With edge-aware filtering and response reformulation, self-similarity responses are computed efficiently in a dense manner.

Interestingly, LSS [26] also uses a max-pooling on the log-polar bins to mitigate the effects of non-rigid deformation. However, the max-pooling in the single self-similarity surface of LSS [26] loses fine-scale matching details as reported in [10]. By contrast, our descriptor employs log-polar spatial pyramid pooling on multiple self-similarity surfaces in order to provide more discriminative representation of self-similarities, thus maintaining matching details as well as providing robustness to non-rigid deformations.

4.1.3 Non-linear Mapping and Normalization

The feature responses are passed through non-linear mapping and normalization to mitigate the effects of outliers. With the accumulated correlation responses $g_i^{\text{ssc}}(l)$, the SSC descriptor $\mathcal{D}_i^{\text{ssc}} = \bigcup_l d_i^{\text{ssc}}(l)$ is computed for $l \in \{1, ..., L^{\text{ssc}}\}$ through a non-linear mapping:

$$d_i^{\text{ssc}}(l) = \exp(-(1 - |g_i^{\text{ssc}}(l)|)/\sigma_c). \tag{4}$$

The features obtained from the SSC descriptor are of size $L^{\text{ssc}} = K \times N_{\mathcal{SB}}$. Finally, $d_i^{\text{ssc}}(l)$ for each pixel i is normalized with an L-2 norm for all l.

4.2 Efficient Computation for Dense Description

The most time-consuming part of SSC is in constructing self-correlation surfaces for all r and j, requiring $K \times M_{\mathcal{R}}^2$ computations of (2) at each pixel i where $M_{\mathcal{R}} \times M_{\mathcal{R}}$ is the size of a local support window $\mathcal{R}$. Straightforward computation of a weighted summation using ω in (2) would require considerable processing with a computational complexity of $O(IM_{\mathcal{F}}^2 K M_{\mathcal{R}}^2)$, where $I = H \times W$ represents the size of an image (height H and width W) and $M_{\mathcal{F}} \times M_{\mathcal{F}}$ is the size of a patch $\mathcal{F}$. To expedite processing, we pre-compute the self-correlation surfaces within a larger local support window, with acceleration via fast edge-aware filtering [55], [56].

First of all, we compute $\mathcal{C}(r, j)$ efficiently by rearranging all sampling patterns (r, j) into reference-biased pairs $(i, h) = (i, i + r - j)$. Similar to DASC [10], $\mathcal{C}(i, h)$ can be expressed in an approximate form[1] as

$$\hat{\mathcal{C}}(i,h) = \frac{\sum\limits_{i',h'} \omega_{i,i'}(f_{i'} - G_i)(f_{h'} - G_h^i)}{\sqrt{\sum\limits_{i'} \omega_{i,i'}(f_{i'} - G_i)^2} \sqrt{\sum\limits_{i',h'} \omega_{i,i'}(f_{h'} - G_h^i)^2}}, \tag{5}$$

1. As shown in [10], there exists marginal performance difference between the asymmetric self-correlation measure in (5) and original one in (2).

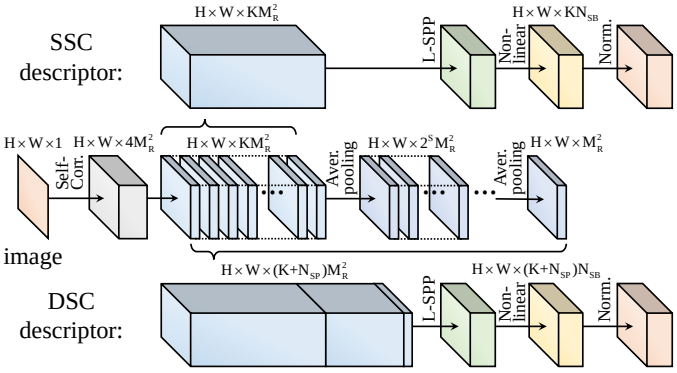


Fig. 6. Visualization of the SSC and DSC descriptors. Our descriptors consist of pyramidal self-correlation computation, log-polar spatial pyramid pooling, non-linear mapping, and normalization.

where $G_h^i = \sum_{i',h'} \omega_{i,i'} f_{h'}$. For faster computation, it can be expressed as follows [10]:

$$\hat{\mathcal{C}}(i,h) = \frac{G_{ih}^i - G_i \cdot G_h^i}{\sqrt{G_{i^2} - (G_i)^2} \cdot \sqrt{G_{h^2}^i - (G_h^i)^2}}, \tag{6}$$

where $G_{ih}^i = \sum_{i',h'} \omega_{i,i'} f_{i'} f_{h'}$, $G_{i^2} = \sum_{i'} \omega_{i,i'} f_{i'}^2$, and $G_{h^2}^i = \sum_{i',h'} \omega_{i,i'} f_{h'}^2$. It can be efficiently computed using any form of fast edge-aware filter [55], [56] with a complexity of $O(IKM_{\mathcal{R}}^2)$. We then simply obtain $\mathcal{C}(r, j)$ from $\hat{\mathcal{C}}(i, h)$ by re-indexing sampling patterns [10].

Though we remove the computational dependency on patch size $M_{\mathcal{F}} \times M_{\mathcal{F}}$, $K \times M_{\mathcal{R}}^2$ computations of (6) are still needed to obtain the self-correlation surfaces, where many sampling pair computations for i and h are repeated. To avoid such redundancy, we first compute a self-correlation surface $\mathcal{C}(i, h)$ for $h \in \mathcal{R}_i^*$ with a doubled local support window $\mathcal{R}_i^*$ of size $2M_{\mathcal{R}} \times 2M_{\mathcal{R}}$. The doubled local support window is used because the minimum support window size for $\mathcal{R}_i^*$ to cover all samples within $\mathcal{R}_i$ is $2M_{\mathcal{R}} \times 2M_{\mathcal{R}}$ as shown in Fig. 5(a). After the self-correlation surface for $\mathcal{R}_i^*$ is computed once over the image domain, $\mathcal{C}(r, j)$ can be extracted through an index mapping process. With this strategy, the computational complexity of constructing self-correlation surfaces becomes $O(I4M_{\mathcal{R}}^2)$, which is smaller than $O(IKM_{\mathcal{R}}^2)$ as $4 \ll K$.

4.3 DSC: Deep Self-Correlation

So far, we have discussed how to build multiple self-similarity surfaces at a single scale and pool the responses. In this section, we extend this idea by encoding self-similar structures at multiple scales. DSC is defined similarly to SSC, except that average pooling is executed before L-SPP (see Fig. 6). Concretely, on multiple self-correlation surfaces, we perform the average pooling using log-polar pyramidal point sets. In comparison to the self-correlations just from a single patch, the aggregation of self-correlation responses is clearly more robust, and it requires only marginal computational overhead over SSC. The strength of such a pyramidal aggregation has also been shown in [46].

Specifically, to build the pyramidal self-correlation surfaces through average pooling, we first define the log-polar pyramidal point sets $\mathcal{SP}_i(v)$ from log-polar point sets $\mathcal{P}_i(l)$, where v indexes all pyramidal levels and all points in each

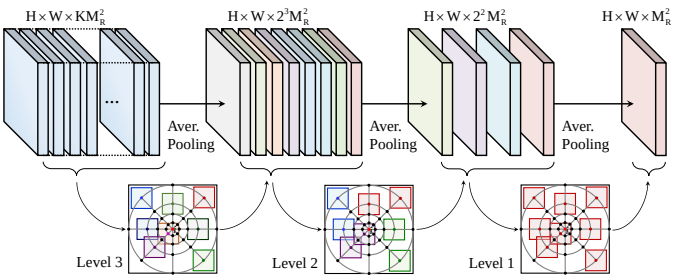


Fig. 7. Visualization of building pyramidal self-correlation surfaces. Multiple self-correlation surfaces are sequentially aggregated using average pooling from the bottom to the top of the log-polar pyramidal point set.

level. $\mathcal{SP}_i$ is defined similarly to $\mathcal{SB}_i$, but on point sets. As shown in Fig. 7, pyramidal self-correlation surfaces are computed by aggregating $\mathcal{C}(r,j)$ for all patches determined on each $\mathcal{SP}(v)$ such that

$$\mathcal{C}(v,j) = \sum_{r \in \mathcal{SP}(v)} \mathcal{C}(r,j)/N_v, \tag{7}$$

which is defined for all v, and N_v is the number of patches within $\mathcal{SP}(v)$. The pyramidal self-correlation surfaces are sequentially aggregated using average pooling from the bottom to the top of the log-polar pyramidal point set. After computing pyramidal self-correlational aggregations, DSC employs L-SPP as well as non-linear mapping and normalization, similar to SSC as presented in Sec. 4.1. A pyramidal self-correlation response is computed as

$$h_i(v,u) = \max_{j \in \mathcal{SB}_i(u)} \mathcal{C}(v,j). \tag{8}$$

We then build a self-correlation response from g_i in (3) and h_i in (8) such that $g_i^{\text{dsc}}(l) = \bigcup_{\{r,v,u\}} \{g_i(r,u), h_i(v,u)\}$ where l indexes over all r, v, and u. Our DSC descriptor $\mathcal{D}_i^{\text{dsc}} = \bigcup_l d_i^{\text{dsc}}(l)$ is then built from $g_i^{\text{dsc}}(l)$ through a non-linear mapping as in (4) for $l \in \{1,...,L^{\text{dsc}}\}$ with $L^{\text{dsc}} = (K+N_{\mathcal{SP}})N_{\mathcal{SB}}$. Finally, $d_i^{\text{dsc}}(l)$ for each pixel i is normalized with an L-2 norm for all l.

4.4 Geometry-invariant DSC

It is known that LSS-based descriptors [10], [24], [26], [61] provide geometric invariance to some extent thanks to its log-polar pooling. However, under more significant geometric variations, existing LSS-based descriptors including DSC do not provide satisfactory performance due to the lack of an explicit module to consider geometric variations. To overcome this issue, we propose geometry-invariant DSC (GI-DSC) that explicitly addresses scale and rotation deformations. The underlying assumption is that geometric deformation fields across cross-modal images can be locally well approximated by a similarity transformation (i.e., translation, rotation, and uniform scale transformation).

4.4.1 Scale-Invariant Multiple Self-Similarities

Existing scale estimation technique as in SIFT [11] is sensitive to cross-modal deformation as exemplified in [30]. We observe that the maximal self-similarities across multiple scales remains consistent with respect to scale changes, and leverage this to provide scale invariance.

Specifically, we first build a Gaussian image pyramid $f^m = f * \varrho_m$, where ϱ_m is the m-th Gaussian kernel for $m = \{1,...,M\}$ and M is the number of Gaussian pyramid levels. For each image pyramid level f^m, we measure the asymmetric self-correlation $\hat{\mathcal{C}}^m(i,h)$, similar to (6), such that

$$\hat{\mathcal{C}}^m(i,h) = \frac{G_{ih}^{i,m} - G_i^m \cdot G_h^{i,m}}{\sqrt{G_{i^2}^m - (G_i^m)^2} \cdot \sqrt{G_{h^2}^{i,m} - (G_h^{i,m})^2}}, \tag{9}$$

where $G_{ih}^{i,m}$, G_i^m, $G_h^{i,m}$, $G_{i^2}^m$, and $G_{h^2}^{i,m}$ are measured for each image pyramid level f_i^m. The scale-invariant self-correlation is then computed by max-pooling as follows:

$$\hat{\mathcal{C}}^{\text{si}}(i,h) = \max_{m \in \{1,...,M\}} \hat{\mathcal{C}}^m(i,h). \tag{10}$$

4.4.2 Orientation Estimation for Rotation Invariance

Similar to scale invariance, rotation invariance can also be achieved by applying multiple orientations to an image. However, such a technique would dramatically increase computational complexity as a function of the product between the number of scales and rotations. Furthermore, our initial experiments indicated that this degrades the localization ability of the descriptor around object boundaries substantially. Fortunately, unlike scale, the orientation field on each pixel can be easily determined from a maximum among orientation histogram weighted by (pre-computed) self-correlations. By transforming the randomly sampled points, the log-polar pyramidal bins, and the log-polar pyramidal points according to the estimated orientation, our descriptor provides rotation invariance on each pixel with only marginal computational overhead.

Specifically, an orientation θ_i of each pixel i is found by constructing a histogram with angles $\angle(i-h)$ for $h \in \mathcal{R}_i^*$ weighted with self-correlations $\hat{\mathcal{C}}(i,h)$ such that

$$l_i^{\text{hist}}(a) = \sum_{h \in \mathcal{H}_i(a)} \hat{\mathcal{C}}(i,h)/N_a, \tag{11}$$

where $\mathcal{H}_i(a) = \{h | h \in \mathcal{R}_i^*, \theta_{a-1} < \angle(i-h) \leq \theta_a\}$ and a quantized angle θ_a for $a \in \{1,...,N_\theta\}$, and N_a is the number of samples in $\mathcal{H}_i(a)$. We then simply choose the main orientation for each pixel corresponding to the most heavily-weighted bin in the histogram, *i.e.*, $\text{argmax}_a l_i^{\text{hist}}(a)$. Moreover, based on the observation that the geometric

Algorithm 1: Deep Self-Correlation (DSC) Descriptor

Input: image f, random samples r
Output: DSC descriptor $\mathcal{D}_i^{\text{dsc}}$
Parameters: number of log-polar pyramidal bins (points) $N_{\mathcal{SB}}(N_{\mathcal{SP}})$
Compute $\hat{\mathcal{C}}(i,h)$ for a doubled support window $\mathcal{R}_i^*$ by using (6).
Compute $\mathcal{C}(r,j)$ from $\hat{\mathcal{C}}(i,h)$ according to the index mapping.
for $v = 1 : N_{\mathcal{SP}}$ **do**
  / **Pyramidal aggregation using average pooling** */*
  Determine a log-polar pyramidal point $\mathcal{SP}_i(v)$.
  Compute $\mathcal{C}(v,j)$ by using average pooling for $\mathcal{SP}_i(v)$ on $\mathcal{C}(r,j)$.
end for
for $u = 1 : N_{\mathcal{SB}}$ **do**
  / **Pyramidal pooling using L-SPP** */*
  Determine a log-polar pyramidal bin $\mathcal{SB}_i(u)$.
  Compute $g_i(r,u)$ and $h_i(v,u)$ by using L-SPP on each $\mathcal{SB}_i(u)$ from $\mathcal{C}(r,j)$ and $\mathcal{C}(v,j)$, respectively.
end for
Build pyramidal self-correlation responses $g_i^{\text{dsc}}(l)$ from $g_i(r,u)$ and $h_i(v,u)$.
Compute a DSC descriptor $\mathcal{D}_i^{\text{dsc}} = \bigcup_l d_i^{\text{dsc}}(l)$, followed by L-2 normalization.

Algorithm 2: Geometry-Invariant DSC (GI-DSC) Descriptor

Input: image f, random samples r
Output: GI-DSC descriptor $\mathcal{D}_i^{\mathrm{gi-dsc}}$
Parameters: number of log-polar pyramidal bins (points) $N_{\mathcal{SB}}(N_{\mathcal{SP}})$
/* ***Scale-invariance*** */
1 : Compute the Gaussian image pyramid $f_i^m = f_i * \varrho_m$.
2 : Compute $\hat{\mathcal{C}}^m(i,h)$ for f_i^m using (9).
3 : Estimate $\hat{\mathcal{C}}^{\mathrm{si}}(i,h)$ using max-pooling as in (10).
/* ***Rotation-invariance*** */
4 : Construct $l_i^{\mathrm{hist}}(a)$ with $\hat{\mathcal{C}}^{\mathrm{si}}(i,h)$ using (11).
5 : Estimate the orientation θ_i for each pixel i from $l_i^{\mathrm{hist}}(a)$.
6 : Filter out the orientation θ_i to provide smooth geometric fields.
7 : Transform r, $\mathcal{SB}_i$, and $\mathcal{SP}_i$ according to θ_i.
8 : Through Step **2-8** in Algorithm 1, compute a GI-DSC descriptor such that $\mathcal{D}_i^{\mathrm{gi-dsc}} = \bigcup_l d_i^{\mathrm{gi-dsc}}(l)$.

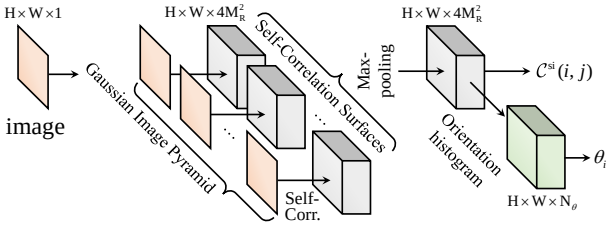


Fig. 8. Visualization of geometry-invariance in GI-DSC. To provide scale invariance, our approach measures multi-scale self-correlation surfaces, and fuses them by max-pooling. Moreover, canonical orientation fields for each pixel are estimated to provide orientation invariance.

deformation fields tend to vary smoothly except at object boundaries [24], [62], the estimated orientation θ_i for each pixel i is regularized using a fast (color-guided) global image filter [63] to correct erroneous rotation fields.

To provide rotation invariance to the DSC descriptor in Sec. 4.3, we transform the randomly sampled points, the log-polar pyramidal bins and the log-polar pyramidal points according to estimated rotation θ_i, and then build the DSC descriptor similarly to Fig. 6. By incorporating both scale- and rotation-invariance within the DSC descriptor, we obtain the GI-DSC descriptor with geometric invariance as well as cross-modal robustness. Fig. 8 illustrates this geometry invariance in the GI-DSC descriptor.

5 Experimental Results and Discussion

5.1 Experimental Settings

In our experiments, we implemented DSC and GI-DSC in Matlab/C++ on an Intel Core i7-3770 CPU at 3.40 GHz with the following fixed parameter settings for all datasets: $\{M_{\mathcal{R}}, M_{\mathcal{F}}, \sigma_c, K, S, N_\rho, N_\phi\} = \{9, 5, 0.5, 32, 3, 4, 16\}$, and $\{N_\theta, M\} = \{32, 4\}$. The feature dimension L of SSC and DSC (or GI-DSC) was fixed to 416 and 585, respectively. We choose the guided filter (GF) for edge-aware filtering in (6), with a smoothness parameter of $\epsilon = 0.03^2$.

In the following, DSC and GI-DSC descriptors were compared to other handcrafted descriptors (SIFT [11], DAISY [12], BRIEF [32], LIOP [35], DaLI [28], LSS [26], SegSIFT [54], SegSID [54], DASC [10], and GI-DASC [30]), recent CNN-based descriptors (MC-CNN [64], VGG[2] [16], FCSS [24], MatchNet (MatchN.) [66], Deep Compare (DeepC.) [67], Deep Descriptor (DeepD.) [21], Learned Invariant Feature Transform (LIFT) [22], L2-Net [23], and

2. In 'VGG', ImageNet pretrained VGG-Net [16] from the bottom conv1 to the conv3-4 layer were used with L_2 normalization [65].

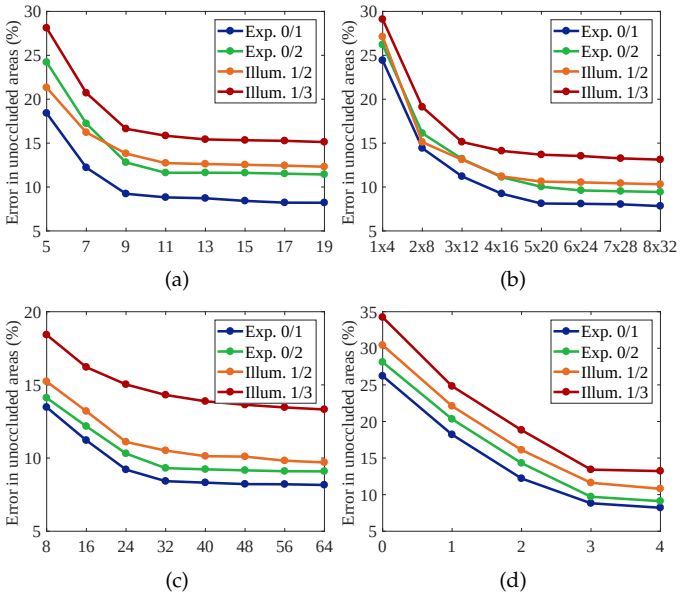


Fig. 9. Component analysis of DSC on the Middlebury benchmark [27] for varying parameter values, such as (a) width (or height) $M_{\mathcal{R}}$ of the local support window, (b) number of log-polar points $N_\rho \times N_\phi$, (c) number of random samples K, and (d) level of log-polar pyramid S. In each experiment, all other parameters are fixed to the initial values.

Quadruplet Network (Q-Net) [37][3]), and area-based similarity measures (ANCC [44] and RSNCC [9]). Furthermore, to evaluate the performance gain by encoding self-similar structures at multiple scales, we compared SSC and DSC. To determine correspondences among the candidates, we used various optimization techniques, such as winner-takes-all (WTA) [64], graph-cut (GC) [68], and SIFT flow (SF) [13], for which the code is publicly available.

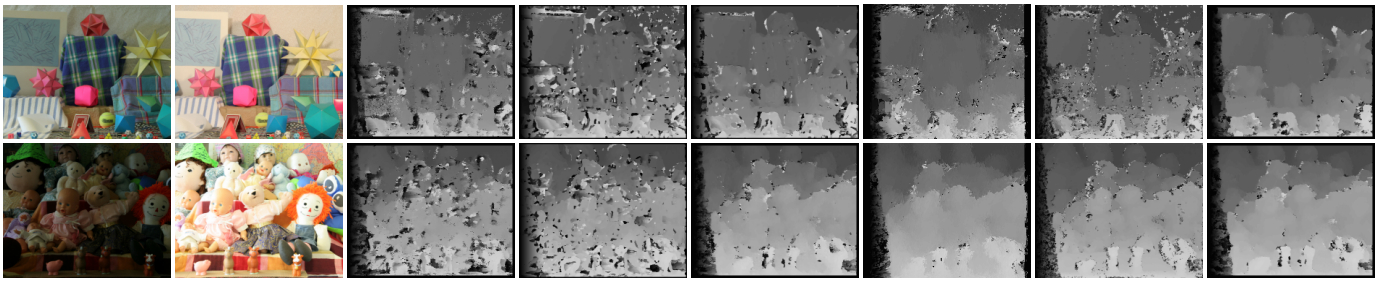
5.2 Ablation Study

Fig. 9 shows the performance of DSC with varying parameter values, including width (or height) $M_{\mathcal{R}}$ of the local support window, number of log-polar points $N_\rho \times N_\phi$, number of random samples K, and levels of the log-polar pyramid S. Fig. 9(c) and (d) demonstrate the effectiveness of self-correlation surfaces and pyramidal structures. For a quantitative analysis, we measured the average bad-pixel error rate in non-occluded areas of disparity maps on the Middlebury benchmark [27]. With a larger support window $M_{\mathcal{R}} \times M_{\mathcal{R}}$, the matching quality improves rapidly until about 9×9. $N_\rho \times N_\phi$ influences the performance of log-polar pooling, which is found to plateau at 4×16. Using a larger number of random samples K yields better performance since DSC encodes more information. The number of log-polar pyramid levels S also affects the amount of encoding. Based on these experiments, we set $K = 32$ and $S = 3$ in consideration of efficiency and robustness.

5.3 Middlebury Stereo Benchmark

We first evaluated SSC and DSC on the Middlebury stereo benchmark [27], which contains illumination and exposure

3. Since MatchN. [66], DeepC. [67], DeepD. [21], LIFT [22], L2-Net [23], and Q-Net [37] were developed for sparse correspondence, sparse descriptors were first built by forward-propagating images through networks and then upsampled.



(a) image 1 (b) image 2 (c) ANCC [44] (d) SIFT [11] (e) DASC [10] (f) FCSS [24] (g) MCCNN [64] (h) DSC

Fig. 10. Comparison of disparity estimates for *Moebius* and *Dolls* image pairs on the Middlebury benchmark [27] across illumination combination '1/3' and exposure combination '0/2', respectively. Compared to other methods, DSC estimates more accurate and edge-preserved disparity maps.

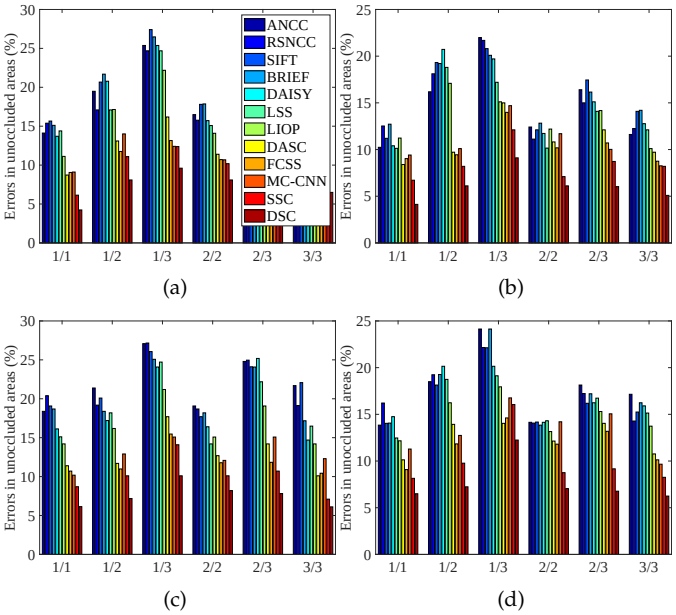
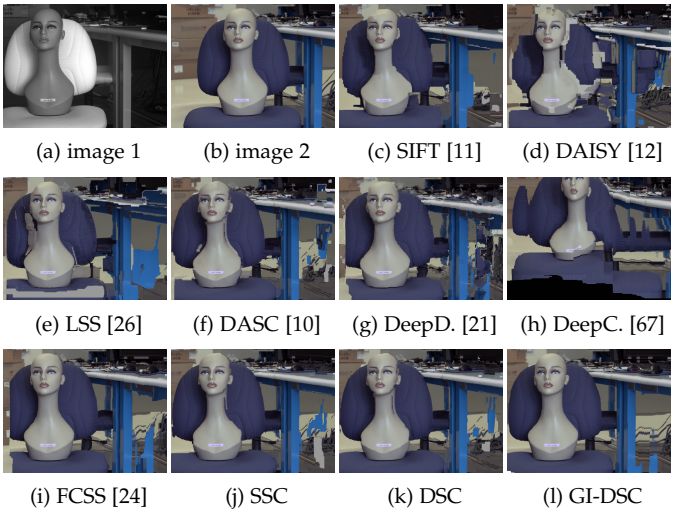


Fig. 11. Average bad-pixel error rate on the Middlebury benchmark [27] with illumination and exposure variations. Optimization was done by GC in (a), (b), and by WTA in (c), (d). SSC and DSC descriptors show the best performance with the lowest error rate.



(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]
(e) LSS [26] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [67]
(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

Fig. 12. Dense correspondence evaluations for RGB-NIR image pairs on cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.

variations. In the experiments, the illumination (exposure) combination '1/3' indicates that two images were captured under the 1^{st} and 3^{rd} illumination (exposure) conditions. For quantitative evaluation, we measured the average bad-pixel error rate in non-occluded areas of disparity maps [27].

Fig. 10 shows the disparity maps estimated under severe illumination and exposure variations with WTA optimization. Fig. 11 displays the average bad-pixel error rates of disparity maps obtained under illumination or exposure variations, with GC [68] and WTA optimization. Note that since the geometric variation across stereo images exists only in the field of translation, GI-DSC was not evaluated in this experiment. Area-based approaches such as ANCC [44] and RSNCC [9] were sensitive to severe radiometric variations, especially when local variations occur frequently. Feature descriptor-based methods such as SIFT [11], DAISY [12], BRIEF [32], LSS [26], and DASC [10] perform better than the area-based approaches, but they also provide limited performance. Although the CNN-based descriptor MC-CNN [64] has shown good results, it exhibits limited performance in cases of severe radiometric variation. Note that since other state-of-the-art stereo matching methods directly estimate disparity maps in an end-to-end manner, they were not evaluated in this experiment. Our DSC achieves the best results both quantitatively and qualitatively. Compared to SSC, the performance of DSC is highly improved, where the performance benefits of leveraging self-similar structures at multiple scales are apparent.

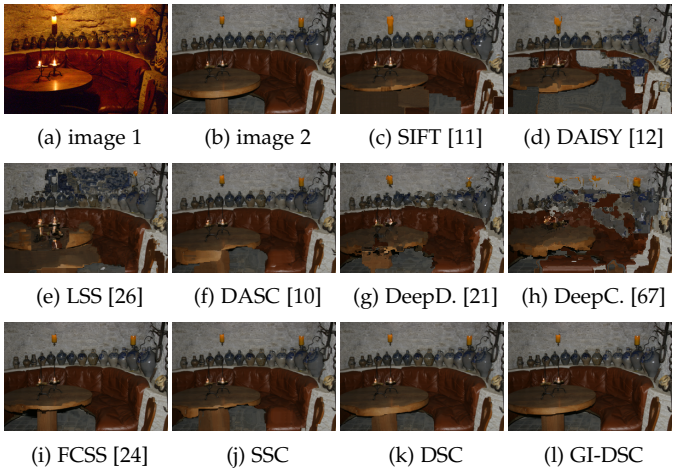
5.4 Cross-modal and Cross-spectral Benchmark

We also evaluated DSC and GI-DSC on the cross-modal and cross-spectral benchmark [10] containing various kinds of image pairs, namely RGB-NIR, flash-noflash, different exposures, and blurred-sharp. Sparse ground-truths for those images were used for error measurement as done in [10].

Fig. 12, Fig. 13, Fig. 14, and Fig. 15 provide qualitative comparisons of the DSC and GI-DSC descriptors to other state-of-the-art approaches for RGB-NIR, flash-noflash, different exposures, and blurred-sharp images, respectively. As already described in the literature [9], gradient-based approaches such as SIFT [11] and DAISY [12] have shown limited performance for RGB-NIR pairs where gradient reversals and inversions frequently appear. BRIEF [32] cannot deal with noisy regions and modality-based appearance differences since it is formulated on pixel differences only. Unlike these approaches, LSS [26] and DASC [10] consider local self-similarities, but LSS suffers from limited discriminative power. DASC also exhibits limited performance due to the sensitivity of patch-wise receptive field pooling. State-of-the-art CNN-based descriptors such as

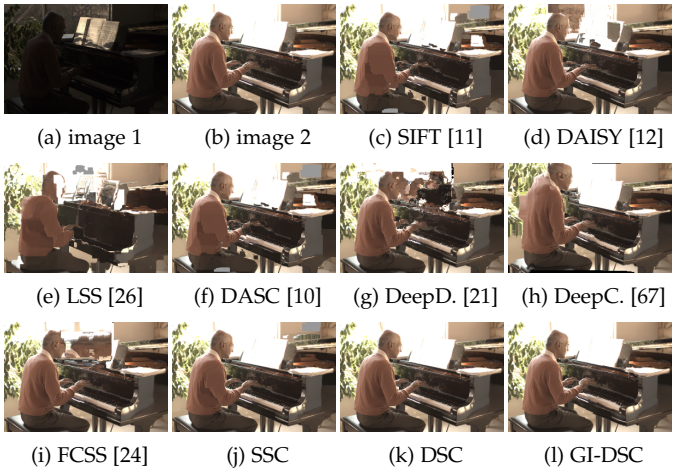
TABLE 1
Average error rates on a cross-modal and cross-spectral benchmark [10]. L2-Net† denotes results of L2-Net [23] with densely sampled windows.

	WTA optimization					SF optimization [13]				
	RGB-NIR	flash-noflash	diff. expo.	blur-sharp	Average	RGB-NIR	flash-noflash	diff. expo.	blur-sharp	Average
ANCC [44]	23.21	20.42	25.19	26.14	23.74	18.45	14.14	11.96	19.24	15.95
RSNCC [9]	27.51	25.12	18.21	27.91	24.69	13.41	15.87	9.15	18.21	14.16
SIFT [11]	24.11	18.72	19.42	27.18	22.36	18.51	11.06	14.87	20.78	16.31
DAISY [12]	27.61	26.30	20.72	27.41	25.51	20.42	10.84	12.71	22.91	16.72
BRIEF [32]	29.14	18.29	17.13	26.43	22.75	17.54	9.21	9.54	19.72	14.00
LSS [26]	27.82	19.18	18.21	26.14	22.84	16.14	11.88	9.11	18.51	13.91
LIOP [35]	24.42	16.42	14.22	20.42	18.87	15.32	11.42	10.22	17.12	13.52
DASC [10]	14.51	13.24	10.32	16.42	13.62	13.42	7.11	7.21	11.21	9.74
MatchN. [66]	19.72	16.54	20.81	27.14	21.05	17.51	10.82	11.84	12.34	13.13
DeepC. [67]	20.71	20.78	16.84	21.84	20.04	17.11	14.21	10.87	11.98	13.54
DeepD. [21]	16.72	17.81	12.72	20.71	16.99	14.87	10.88	12.87	13.93	13.14
Q-Net [37]	10.11	16.75	12.81	22.95	15.66	10.40	17.42	13.92	12.38	13.53
LIFT [22]	14.82	14.32	10.11	17.84	14.27	12.88	10.28	9.77	10.54	10.87
L2-Net [23]	13.79	13.16	9.92	19.11	13.99	11.92	15.22	11.20	11.69	12.51
L2-Net† [23]	12.61	14.22	10.22	20.54	14.40	10.51	14.66	10.90	12.17	12.06
FCSS [24]	11.87	**9.84**	**7.99**	17.64	11.84	12.10	6.28	6.11	10.84	8.83
SSC	**10.12**	10.12	8.22	**14.22**	**10.67**	**9.12**	**6.18**	**5.22**	**9.12**	**7.41**
DSC	**8.12**	**8.22**	**6.72**	**13.28**	**9.09**	**7.62**	**5.12**	**4.72**	**8.01**	**6.37**
GI-DSC	**9.30**	**7.92**	**6.86**	**12.92**	**9.25**	**7.12**	**4.75**	**4.42**	**7.06**	**5.84**



(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]
(e) LSS [26] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [67]
(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

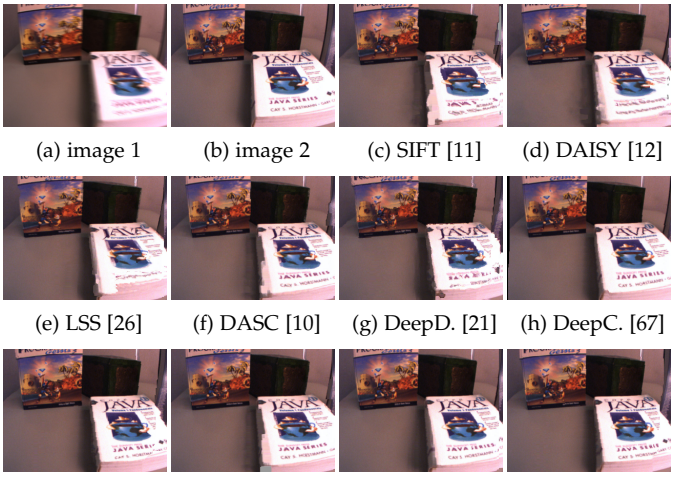
Fig. 13. Dense correspondence evaluations for flash-noflash image pairs on cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.



(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]
(e) LSS [26] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [67]
(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

Fig. 14. Dense correspondence evaluations for different exposure image pairs on a cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.

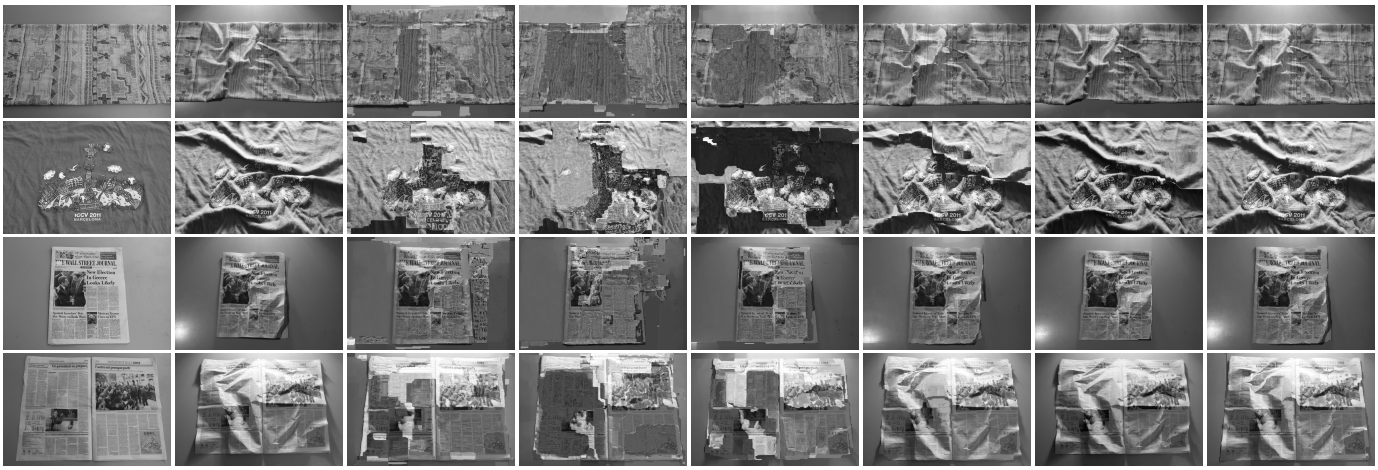
MatchN. [66], DeepC. [67], DeepD. [21], LIFT [22], L2-Net [23], and FCSS [24], pretrained on non-cross-modal



(a) image 1 (b) image 2 (c) SIFT [11] (d) DAISY [12]
(e) LSS [26] (f) DASC [10] (g) DeepD. [21] (h) DeepC. [67]
(i) FCSS [24] (j) SSC (k) DSC (l) GI-DSC

Fig. 15. Dense correspondence evaluations for blurred-sharp image pairs on a cross-modal and cross-spectral benchmark [10]. The source images were warped to the target images using correspondences.

image pairs, cannot provide reliable correspondence estimation performance on cross-modal matching. Even though those methods have shown high robustness to photometric variations, they provide limited precision in localization. Moreover, large-scale training datasets are lacking for learning those descriptors. Q-Net [37] trained on the RGB-NIR dataset [1] has shown limited generalization ability to the appearance variations of various modalities such as flash-noflash, different exposures, and blurred-sharp. Compared to those methods, DSC displays better correspondence estimation. We also performed a quantitative evaluation with results listed in Table 1, which also clearly demonstrates the effectiveness of DSC. Note that the geometric variation across images provided from the cross-modal and cross-spectral benchmark [10] is not substantial, and thus it is relatively difficult to show the effectiveness of GI-DSC in terms of handling geometry variation. Nevertheless, GI-DSC demonstrates improved performance over DSC.



(a) image 1 (b) image 2 (c) DAISY [12] (d) BRIEF [32] (e) LSS [26] (f) DaLI [28] (g) FCSS [24] (g) DSC

Fig. 16. Dense correspondence evaluations for images with different illumination conditions and non-rigid image deformations [28]. The source images were warped to the target images using correspondences.

TABLE 2
Average error rates on the DaLI benchmark [28].

Methods	deform.	illum.	deform./ illum.	Average
DAISY [12]	43.98	42.72	43.42	43.37
BRIEF [32]	41.51	37.14	41.35	40.00
LSS [26]	40.81	39.54	40.11	40.12
LIOP [35]	28.72	31.72	30.21	30.22
DaLI [28]	27.12	27.31	27.99	27.47
DASC [10]	26.21	24.83	27.51	26.18
VGG [16]	25.72	23.41	22.51	23.88
LIFT [22]	27.42	27.11	29.28	27.94
L2-Net [23]	26.34	25.74	26.84	26.31
FCSS [24]	**22.18**	24.72	**19.72**	**22.21**
SSC	23.42	**22.21**	24.17	23.27
DSC	**20.14**	**20.72**	**21.87**	**20.91**
GI-DSC	**18.47**	**16.25**	**18.24**	**17.65**

TABLE 3
Average error rates on the tri-modal human benchmark [29].

	RGB-depth		RGB-thermal		depth-thermal	
	LTA	IoU	LTA	IoU	LTA	IoU
DAISY [12]	45.51	0.41	36.31	0.44	53.21	0.52
BRIEF [32]	46.22	0.46	48.11	0.41	57.22	0.53
LSS [26]	49.27	0.52	49.38	0.42	51.87	0.42
LIOP [35]	41.75	0.37	48.27	0.36	50.78	0.39
DaLI [28]	40.99	0.39	48.72	0.43	53.95	0.50
DASC [10]	36.72	0.36	38.27	0.39	43.72	0.41
VGG [16]	33.16	0.39	38.11	0.42	46.72	0.38
LIFT [22]	38.72	0.47	43.51	0.49	50.84	0.53
L2-Net [23]	36.27	0.41	38.84	0.38	42.54	0.47
FCSS [24]	30.82	0.31	**29.71**	**0.30**	**39.78**	**0.34**
SSC	**30.11**	**0.29**	30.87	0.31	42.81	0.36
DSC	**26.19**	**0.24**	**29.38**	**0.27**	**36.22**	**0.27**
GI-DSC	**22.63**	**0.19**	**27.42**	**0.24**	**30.82**	**0.21**

5.5 DaLI Benchmark

We also evaluated the DSC and GI-DSC descriptors on a publicly available dataset featuring challenging non-rigid deformations and severe illumination changes [28]. Fig. 16 shows dense correspondence estimates for this benchmark [28]. A quantitative evaluation is given in Table 2 using ground-truth feature points sparsely extracted for each image. As expected, conventional gradient-based and intensity comparison-based feature descriptors, including SIFT [11], DAISY [12], and BRIEF [32], are relatively less effective on such images. LSS [26] and DASC [10] exhibit relatively high performance for illumination changes, but perform less well on non-rigid geometric deformations. LIOP [35] provided robustness to radiometric variations, but is sensitive to non-rigid deformations. Although DaLI [28] estimated robust correspondences, it requires considerable computation for dense matching. DSC offers greater discriminative power as well as more robustness to non-rigid deformations in comparison to the state-of-the-art cross-modality descriptors. State-of-the-art deep CNN-based methods such as FCSS [24] also show strong performance but require considerable training time and a large number of training samples. By using explicit geometric estimation modules, GI-DSC presents state-of-the-art performance for non-rigid deformations.

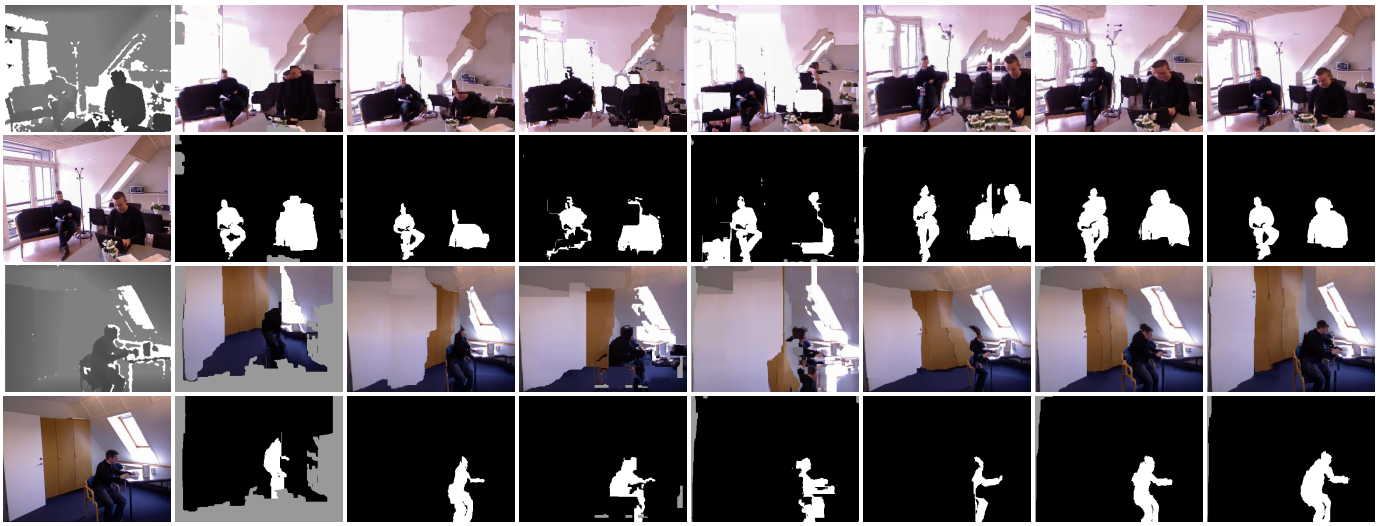
5.6 Tri-modal Human Benchmark

We additionally evaluated our descriptors on the tri-modal human body segmentation dataset [29] which includes RGB-Depth-FIR pairs. The dataset contains 11,537 frames divided into three indoor scenes and among them, 5,724 frames have human body annotations. To quantitatively measure the estimated correspondence quality, we use the label transfer accuracy (LTA) [13], [30] and intersection over union (IoU) metrics [14], [24] with ground-truth annotation maps, a practical alternative when no ground-truth correspondence is available.

Fig. 17 and Fig. 18 display qualitative comparisons for RGB-Depth and RGB-FIR pairs, respectively. Table 3 presents a quantitative evaluation in terms of LTA and IoU. In comparison to the experiments of previous sections, this experiment uses RGB-Depth, RGB-FIR, and Depth-FIR pairs with more severe cross-modal variations. Similar to the previous experiments, conventional handcrafted descriptors such as DASC [10] show limited performance. Although state-of-the-art CNN-based methods produce improvements, they cannot deal with non-rigid deformations or severe appearance variations across cross-modal images.

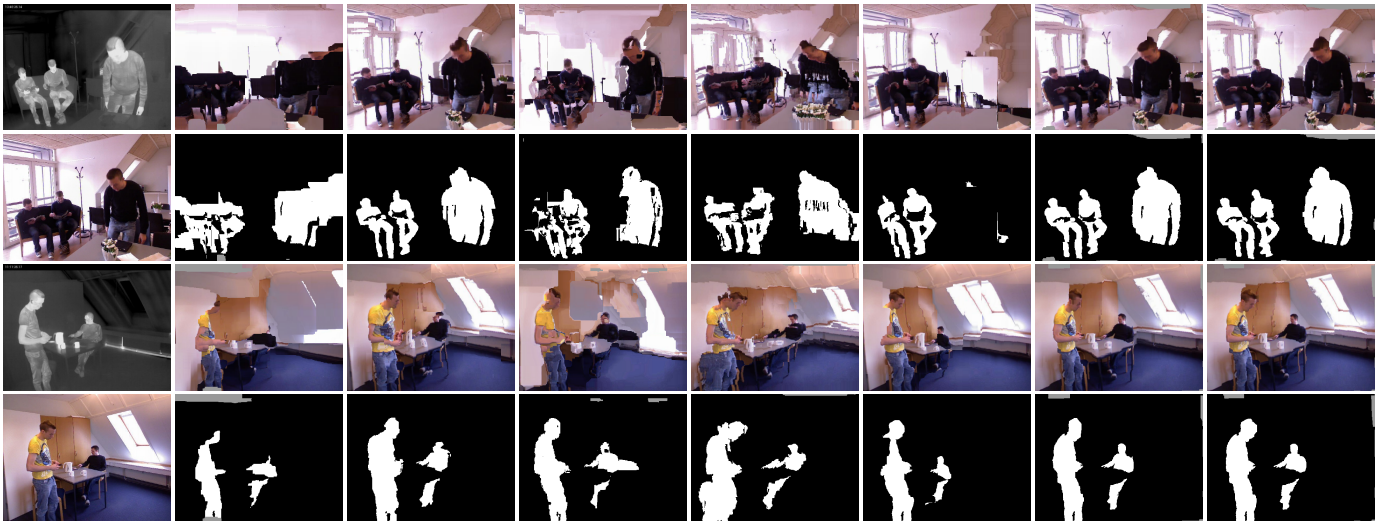
5.7 DIML Cross-modal Benchmark

We further evaluated the DSC and GI-DSC descriptors on the DIML cross-modal benchmark [30] with both photomet-



(a) Image pairs (b) SIFT [11] (c) DASC [10] (d) LSS [26] (e) VGG [16] (f) FCSS [24] (g) DSC (h) GI-DSC

Fig. 17. Qualitative comparisons on the RGB-depth human benchmark [29]. The results consist of warped color images and warped ground-truth human annotations.



(a) Image pairs (b) SIFT (c) DASC [10] (d) LSS [26] (e) LIFT [22] (f) FCSS [24] (g) DSC (h) GI-DSC

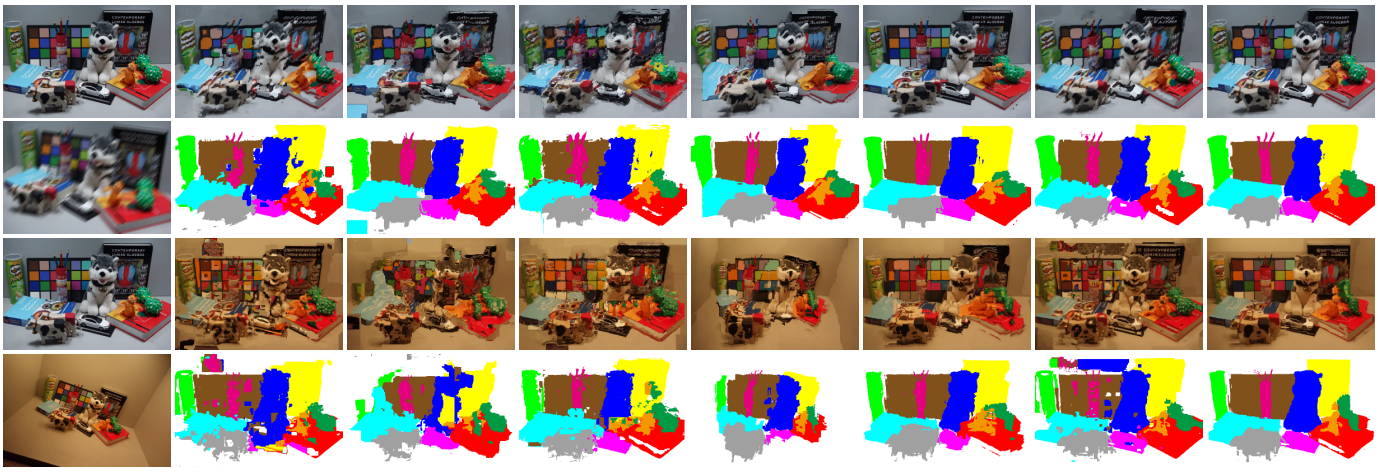
Fig. 18. Qualitative comparisons on the RGB-thermal human benchmark [29]. The results consist of warped color images and warped ground-truth human annotations.

ric and geometric variations. In the benchmark, 10 geometry image sets were captured with geometric variations that arise from a combination of viewpoint, scale, and rotation differences, and each image set consists of images taken under 5 different photometric variation pairs including illumination, exposure, flash-noflash, blur, and noise. The DIML cross-modal benchmark thus consists of 100 images, of size 1200×800. For quantitative evaluation, we used LTA and IoU, similar to Sec. 5.6. Note that different from LTA, IoU isolates the matching quality for foreground objects, separate from irrelevant background pixels.

In Fig. 19 and Fig. 20, we followed the experimental configuration in [30], where for an image from a reference geometry set, we estimate visual correspondence maps with images from other geometry sets, and measure LTA. Furthermore, visual correspondence maps are estimated for each photometric pair. In addition, in Table 4, we measured the average LTA and IoU for all possible photometric variation pairs with fixed geometry (denoted by *photometry*) and all possible geometric variation pairs with fixed photometry (denoted by *geometry*), respectively, and all possible geometric and photometric variations (denoted by *all*) [4]

As expected, conventional gradient-based and intensity comparison-based feature descriptors, including SIFT [11], DAISY [12], and BRIEF [32], provided weaker correspondence performance. LSS [26] and DASC [10] exhibited relatively high performance for illumination changes, but are limited on non-rigid deformations. LIOP [35] provided robustness to radiometric variations, but is sensitive to non-rigid deformations. Although DaLI [28] yielded robust correspondences, it requires considerable computation

4. Specifically, for *photometry* results, we sampled image pairs among 5 photometric variations for 10 geometric variations, i.e., the number of image pairs is ${}^5C_2 \times 10$. For *geometry* results, we sampled image pairs among 10 geometric variations for 5 geometric variations, i.e., the number of image pairs is ${}^{10}C_2 \times 5$. For *all* results, we sampled all possible combinations, i.e., the number of image pairs is ${}^{50}C_2$.



(a) Image pairs (b) DAISY (c) Seg-SIFT [54] (d) Seg-SID [54] (e) DASC [10] (f) GI-DASC [30] (g) DSC (h) GI-DSC

Fig. 19. Qualitative comparisons on the DIML cross-modal benchmark [69]. The results consist of warped color images and warped ground-truth annotations.

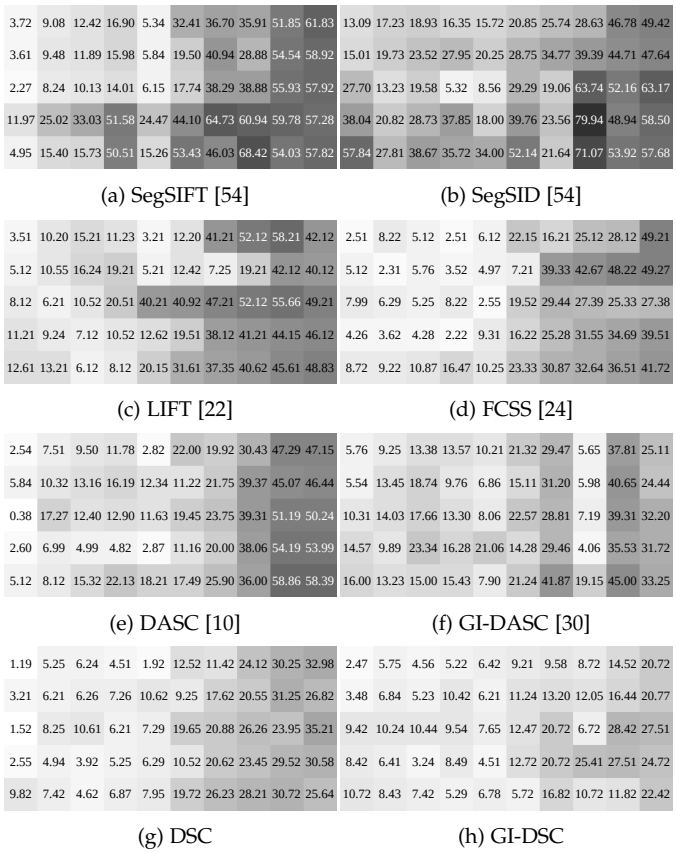


Fig. 20. Quantitative comparisons on the DIML cross-modal benchmark [69]. Each result represents the LTA for geometric (x-axis) and photometric (y-axis) variations, respectively.

for dense matching. State-of-the-art CNN-based descriptors such as LIFT [22] and FCSS [24] cannot deal with photometric and geometric variations simultaneously, resulting in limited performance. DSC offers greater discriminative power as well as more robustness to non-rigid deformation in comparison to the state-of-the-art cross-modality descriptors, but it remains vulnerable to severe geometric variations. Unlike these, GI-DSC shows robustness to both photometric and geometric variations.

TABLE 4
Average error rates on the DIML cross-modal benchmark [69].

	photometry		geometry		all	
	LTA	IoU	LTA	IoU	LTA	IoU
DAISY [12]	36.42	0.42	48.42	0.51	48.25	0.52
BRIEF [32]	40.51	0.50	47.21	0.54	49.02	0.57
LSS [26]	38.51	0.42	47.80	0.43	47.22	0.48
LIOP [35]	26.71	0.36	52.03	0.41	42.22	0.49
DaLI [28]	34.71	0.34	49.82	0.39	52.11	0.42
Seg-SIFT [54]	28.99	0.38	39.02	0.33	46.42	0.48
Seg-SID [54]	32.76	0.29	51.22	0.40	52.68	0.45
DASC [10]	20.41	0.31	30.81	0.33	32.53	0.38
GI-DASC [24]	21.92	0.32	**21.84**	0.26	27.11	0.31
VGG [16]	22.07	0.29	41.11	0.27	39.62	0.30
LIFT [22]	23.11	0.30	42.02	0.31	35.00	0.38
L2-Net [23]	26.75	0.35	38.74	0.45	36.92	0.42
FCSS [24]	**18.72**	**0.27**	31.80	**0.24**	**30.11**	**0.29**
SSC	19.78	0.29	31.71	0.28	31.62	0.30
DSC	**16.72**	**0.24**	**26.11**	**0.25**	**24.12**	**0.27**
GI-DSC	**14.70**	**0.19**	**16.27**	**0.20**	**19.84**	**0.23**

5.8 Computational Speed

In Fig. 21, we compare the computation speed of DSC and GI-DSC to the state-of-the-art descriptors. Although deep CNN-based descriptors such as LIFT [22] and FCSS [24] are efficient at testing time compared to handcrafted descriptors such as DaLI [28], SIFT [11], and LSS [26], they entail a large computational burden at training time and require a large number of training samples. Compared to the brute-force implementation of DSC, the efficient implementation of DSC greatly reduces computation time. Moreover, compared to DSC, GI-DSC needs only marginal additional computation while providing high geometric invariance. Even though the DSC and GI-DSC descriptors need more computation compared to some previous dense descriptors, they provide significantly improved matching performance as described previously and are training-free.

6 Conclusion

In this paper, we showed recent state-of-the-art CNN-based descriptors even cannot provide satisfactory performances for establishing dense correspondences between images taken under different imaging modalities, and further proposed DSC and GI-DSC descriptors as alternatives. Their

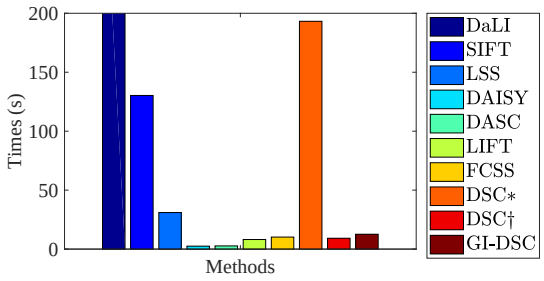


Fig. 21. Computation speed of DSC and GI-DSC descriptors and other state-of-the-art descriptors. The brute-force and efficient implementations of DSC are denoted by * and †, respectively.

high performance in comparison to state-of-the-art descriptors can be attributed to greater robustness to non-rigid deformations because of their effective pooling scheme, and more importantly their heightened discriminative power from a more comprehensive representation of self-similar structure and their formulation in a pyramidal manner. Over an extensive set of experiments that cover a broad range of cross-modal differences, DSC and GI-DSC were validated by their higher performance in comparison to existing handcrafted and deep CNN-based descriptors. Thanks to their robustness to non-rigid deformations and high discriminative power, DSC and GI-DSC can potentially be used to benefit object detection and semantic segmentation in future work.

ACKNOWLEDGMENTS

This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (NRF-2017M3C4A7069370). This work of S. Kim was supported (in part) by the Yonsei University Research Fund (Yonsei Frontier Lab. Young Researcher Supporting Program) of 2018. The work of D. Min was supported by R&D program for Advanced Integrated-intelligence for IDentification (AIID) through the National Research Foundation of Korea(NRF) funded by Ministry of Science and ICT (2018M3E3A1057303).

REFERENCES

[1] M. Brown and S. Susstrunk, "Multispectral sift for scene category recognition," *In: CVPR*, 2011.
[2] Q. Yan, X. Shen, L. Xu, and S. Zhuo, "Cross-field joint image restoration via scale map," *In: ICCV*, 2013.
[3] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," *In: CVPR*, 2015.
[4] D. Krishnan and R. Fergus, "Dark flash photography," *In: SIGGRAPH*, 2009.
[5] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes," *In: SIGGRAPH*, 2012.
[6] Y. HaCohen, E. Shechtman, and E. Lishchinski, "Deblurring by example using dense correspondence," *In: ICCV*, 2013.
[7] H. Lee and K. Lee, "Dense 3d reconstruction from severely blurred images using a single moving camera," *In: CVPR*, 2013.
[8] G. Petschnigg, M. Agrawals, and H. Hoppe, "Digital photography with flash and no-flash image pairs," *In: SIGGRAPH*, 2004.
[9] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multi-spectral registration for natural images," *In: ECCV*, 2014.
[10] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "Dasc: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," *In: CVPR*, 2015.
[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
[12] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. PAMI*, vol. 32, no. 5, pp. 815–830, 2010.
[13] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Trans. PAMI*, vol. 33, no. 5, pp. 815–830, 2011.
[14] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," *In: CVPR*, 2013.
[15] P. Pinggera, T. Breckon, and H. Bischof, "On cross-spectral stereo matching using dense gradient features," *In: BMVC*, 2012.
[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *In: ICLR*, 2015.
[17] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. PAMI*, vol. 36, no. 8, pp. 1573–1585, 2014.
[18] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional acitivation features," *In: ECCV*, 2014.
[19] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: A comparison to sift," *arXiv:1405.5769*, 2014.
[20] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *In: ICML*, 2014.
[21] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," *In: ICCV*, 2015.
[22] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," *In: ECCV*, 2016.
[23] Y. Tian, B. Fan, and F. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," *In: CVPR*, 2017.
[24] S. Kim, D. Min, B. Ham, S. Jeon, S. Lin, and K. Sohn, "Fcss: Fully convolutional self-similarity for dense semantic correspondence," *In: CVPR*, 2017.
[25] J. Dong and S. Soatto, "Domain-size pooling in local descriptors: Dsp-sift," *In: CVPR*, 2015.
[26] E. Schechtman and M. Irani, "Matching local self-similarities across images and videos," *In: CVPR*, 2007.
[27] D. Scharstein and R. Szeliski, "A taxanomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-3, pp. 7–42, 2002.
[28] E. Simo-Serra, C. Torras, and F. Moreno-Noguer, "Dali: Deformation and light invariant descriptor," *IJCV*, vol. 115, no. 2, pp. 136–154, 2015.
[29] C. Palmero, A. Clapes, C. Bahnsen, and A. Mogelmose, "Multimodal rgb-depth-thermal human body segmentation," *IJCV*, vol. 118, no. 2, pp. 217–239, 2016.
[30] 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 Trans. PAMI*, vol. 39, no. 9, pp. 1712–1729, 2017.
[31] S. Kim, D. Min, S. Lin, and K. Sohn, "Deep self-correlation descriptor for dense cross-modal correspondence," *In: ECCV*, 2016.
[32] M. Calonder, "Brief : Computing a local binary descriptor very fast," *IEEE Trans. PAMI*, vol. 34, no. 7, pp. 1281–1298, 2011.
[33] T. Trzcinski, M. Christoudias, and V. Lepetit, "Learning image descriptor with boosting," *IEEE Trans. PAMI*, vol. 37, no. 3, pp. 597–610, 2015.
[34] S. Saleem and R. Sablatnig, "A robust sift descriptor for multispectral images," *IEEE SPL*, vol. 21, no. 4, pp. 400–403, 2014.
[35] Z. Wang, B. Fan, and F. Wu, "Local intensity order pattern for feature description," *In: ICCV*, 2011.
[36] 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: CVPR Workshop*, 2016.
[37] C. A. Aguilera, A. D. Sappa, C. Aguilera, and R. Toledo, "Cross-spectral local descriptors via quadruplet network," *In: Sensors*, vol. 17, no. 4, 2017.
[38] P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, V. Gleeson, S. Brady, and A. Schnabel, "Mind: Modality indepdent neighbourhood descriptor for multi-modal deformable registration," *MIA*, vol. 16, no. 3, pp. 1423–1435, 2012.

[39] A. Torabi and G. Bilodeau, "Local self-similarity-based registration of human rois in pairs of stereo thermal-visible videos," *PR*, vol. 46, no. 2, pp. 578–589, 2013.
[40] Y. Ye and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *JPRS*, vol. 90, no. 7, pp. 83–95, 2014.
[41] J. Pluim, J. Maintz, and M. Viergever, "Mutual information based registration of medical images: A survey," *IEEE Trans. MI*, vol. 22, no. 8, pp. 986–1004, 2003.
[42] Y. Heo, K. Lee, and S. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," *IEEE Trans. PAMI*, vol. 35, no. 5, pp. 1094–1106, 2013.
[43] J. Xu, Q. Yang, J. Tang, and Z. Feng, "Linear time illumination invariant stereo matching," *IJCV*, 2016.
[44] Y. Heo, K. Lee, and S. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Trans. PAMI*, vol. 33, no. 4, pp. 807–822, 2011.
[45] M. Irani and P. Anandan, "Robust multi-sensor image alignment," *In: ICCV*, 1998.
[46] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Deepflow: Large displacement optical flow with deep matching," *In: ICCV*, 2013.
[47] T. Hassner, V. Mayzels, and L. Zelnik-Manor, "On sifts and their scales," *In: CVPR*, 2012.
[48] W. Qiu, X. Wang, X. Bai, A. Yuille, and Z. Tu, "Scale-space sift flow," *In: WACV*, 2014.
[49] J. Hur, H. Lim, C. Park, and S. C. Ahn, "Generalized deformable spatial pyramid: Geometry-preserving dense correspondence estimation," *In: CVPR*, 2015.
[50] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," *In: ECCV*, 2010.
[51] H. Yang, W. Lin, and J. Lu, "Daisy filter flow: A generalized discrete approach to dense correspondences," *In: CVPR*, 2014.
[52] J. Lu, H. Yang, D. Min, and M. N. Do, "Patchmatch filter: Efficient edge-aware filtering meets randomized search for fast correspondence field estimation," *In: CVPR*, 2013.
[53] I. Kokkinos and A. Yuille, "Scale invariance without scale selection," *In: CVPR*, 2008.
[54] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. M. Noguer, "Dense segmentation-aware descriptors," *In: CVPR*, 2013.
[55] E. Gastal and M. Oliveira, "Domain transform for edge-aware image and video processing," *In: SIGGRAPH*, 2011.
[56] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. PAMI*, vol. 35, no. 6, pp. 1397–1409, 2013.
[57] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *In: CVPR*, 2005.
[58] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *In: CVPR*, 2006.
[59] L. Seidenari, G. Serra, A. D. Bagdanov, and A. D. Bimbo, "Local pyramidal descriptors for image recognition," *IEEE Trans. PAMI*, vol. 36, no. 5, pp. 1033–1040, 2014.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. PAMI*, vol. 37, no. 9, pp. 1904–1916, 2015.
[61] K. Chatfield, J. Philbin, and A. Zisserman, "Efficient retrieval of deformable shape classes using local self-similarities," *In: ICCV Workshop*, 2009.
[62] M. Tau and T. Hassner, "Dense correspondences across scenes and scales," *IEEE Trans. PAMI*, vol. 38, no. 5, pp. 875–888, 2016.
[63] D. Min, S. Choi, J. Lu, B. Ham, K. Sohn, and M. N. Do, "Fast global image smoothing based on weighted least squares," *IEEE Trans. IP*, vol. 23, no. 12, pp. 5638–5653, 2014.
[64] J. Zbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *Journal of Machine Learning Research*, vol. 17, pp. 1–32, 2016.
[65] H. O. Song, Y. Xiang, S. Jegelk, and S. Savarese, "Deep metric learing via lifted structured feature embedding," *In: CVPR*, 2016.
[66] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," *In: CVPR*, 2015.
[67] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," *In: CVPR*, 2015.
[68] Y. Boykov, O. Yeksler, and R. Zabih, "Fast approximation enermgy minimization via graph cuts," *IEEE Trans. PAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.
[69] online., http://diml.yonsei.ac.kr/∼srkim/DASC/.

Seungryong Kim received the B.S. and Ph.D. degrees in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 2012 and 2018, respectively. From 2018 to 2019, he was Post-Doctoral Researcher in Yonsei University, Seoul, Korea. Since 2019, he has been Post-Doctoral Researcher in School of Computer and Communication Sciences at École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. His current research interests include 2D/3D computer vision, computational photography, and machine learning.

Dongbo Min received the B.S., M.S., and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2003, 2005, and 2009, respectively. From 2009 to 2010, he was a post-doctoral researcher with Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. From 2010 to 2015, he was with the Advanced Digital Sciences Center, Singapore. From 2015 to 2018, he was an assistant professor with the Department of Computer Science and Engineering, Chungnam National University, Daejeon, South Korea. Since 2018, he has been an assistant professor with the Department of Computer Science and Engineering, Ewha Womans University, Seoul. His current research interests include computer vision, deep learning, video processing, and continuous/discrete optimization.

Stephen Lin received the B.S.E. degree in electrical engineering from Princeton University, NJ, and the Ph.D. degree in computer science and engineering from the University of Michigan, Ann Arbor. He is a Principal Researcher with the Visual Computing group, Microsoft Research Asia. His research interests include computer vision, image processing, and computer graphics. He served as a Program Co-Chair of the International Conference on Computer Vision 2011 and the Pacific-Rim Symposium on Image and Video Technology 2009.

Kwanghoon Sohn received the B.E. degree in electronic engineering from Yonsei University, Seoul, Korea, in 1983, the M.S.E.E. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University, Raleigh, NC, USA, in 1992. He was a Senior Member of the Research engineer with the Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, Korea, from 1992 to 1993, and a Post-Doctoral Fellow with the MRI Center, Medical School of Georgetown University, Washington, DC, USA, in 1994. He was a Visiting Professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently an Underwood Distinguished Professor with the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3D image processing and computer vision.