DASC: Robust Dense Descriptor for Multi-Modal and Multi-Spectral Correspondence Estimation

Seungryong Kim, *Student Member, IEEE*, Dongbo Min, *Senior Member, IEEE*,
Bumsub Ham, *Member, IEEE*, Minh N. Do, *Fellow, IEEE*, and Kwanghoon Sohn, *Senior Member, IEEE*

Abstract—Establishing dense correspondences between multiple images is a fundamental task in many applications. However, finding a reliable correspondence between multi-modal or multi-spectral images still remains unsolved due to their challenging photometric and geometric variations. In this paper, we propose a novel dense descriptor, called dense adaptive self-correlation (DASC), to estimate dense multi-modal and multi-spectral correspondences. Based on an observation that self-similarity existing within images is robust to imaging modality variations, we define the descriptor with a series of an adaptive self-correlation similarity measure between patches sampled by a randomized receptive field pooling, in which a sampling pattern is obtained using a discriminative learning. The computational redundancy of dense descriptors is dramatically reduced by applying fast edge-aware filtering. Furthermore, in order to address geometric variations including scale and rotation, we propose a geometry-invariant DASC (GI-DASC) descriptor that effectively leverages the DASC through a superpixel-based representation. For a quantitative evaluation of the GI-DASC, we build a novel multi-modal benchmark as varying photometric and geometric conditions. Experimental results demonstrate the outstanding performance of the DASC and GI-DASC in many cases of dense multi-modal and multi-spectral correspondences.

Index Terms—Dense correspondence, descriptor, multi-spectral, multi-modal, edge-aware filtering

✦

1 INTRODUCTION

RECENTLY, many computer vision and computational photography problems have been reformulated to overcome their inherent limitations by leveraging multi-modal and multi-spectral images. Typical examples of other imaging modalities include near-infrared (NIR) image [1], [2], depth image [3], and dark flash image [4]. More broadly, flash and no-flash images [5], blurred images [6], [7], and images taken under different radiometric conditions [8] can also be considered as multi-modal [9].

Establishing dense visual correspondences for multi-modal and multi-spectral images is a key enabler for realizing such tasks. In general, the performance of correspondence algorithms relies primarily on two components: appearance descriptor and optimization scheme. Traditional dense correspondence methods for estimating depth [10] or optical flow [11], [12] fields, in which input images are acquired in a similar imaging condition, have been dramatically advanced in recent studies. To define a matching fidelity term, they typically assume that multiple images share a similar visual pattern, e.g., color, gradient, and structural similarity. However, when it comes to multi-spectral and multi-modal images, such properties do not hold as shown in Fig. 1, and thus conventional descriptors or similarity measures often fail to capture reliable matching evidence. This leads to a poor matching quality as shown in Fig. 2. Furthermore, substantial geometric variations, which often appear in images captured under wide-baseline conditions, make the matching task even more challenging. Although employing a powerful optimization technique could help estimate a reliable solution with a spatial context [14], [15], [16], an optimizer itself cannot address an inherent limitation without suitable matching descriptors [17].

Our method starts from an observation that a local internal layout of self-similarities is less sensitive to photometric distortions, even when an intensity distribution of an anatomical structure is not maintained across different imaging modalities [18]. That is, the local self-similarity (LSS) descriptor would be beneficial to overcoming inherent limitations of existing descriptors in establishing correspondences between multi-modal or multi-spectral images. Several approaches based on the LSS have been presented for multi-modal and multi-spectral image registration [19], [20], but they do not scale well to estimating dense correspondences for multi-modal and multi-spectral images, and their matching performance is still poor.

In this paper, we propose a novel local descriptor, called dense adaptive self-correlation (DASC), designed for establishing dense multi-modal and multi-spectral correspondences. It is defined with a series of patch-wise similarities within a local support window. The similarity is computed with an adaptive self-correlation (ASC) measure, which

- *S. Kim, B. Ham, and K. Sohn are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea. E-mail: {srkim89, mimo, khsohn}@yonsei.ac.kr.*
- *D. Min is with the Department of Computer Science and Engineering, Chungnam National University, Daejeon 305-764, Korea. E-mail: dbmin@cnu.ac.kr.*
- *M.N. Do is with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801. E-mail: minhdo@illinois.edu.*

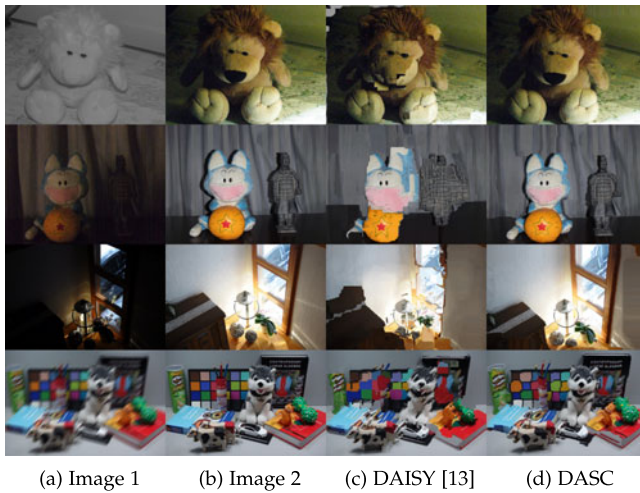| (a) Image 1 | (b) Image 2 | (c) DAISY [13] | (d) DASC |

Fig. 1. Some challenging multi-modal and multi-spectral images such as (from top to bottom) RGB-NIR, flash-noflash images, two images with different exposures, and blur-sharp images. The images in the third and fourth column are the results obtained by warping images in the second column to images in the first column with dense correspondence maps estimated by using DAISY [13] and our DASC descriptor, respectively.

(a) RGB image   (b) NIR image

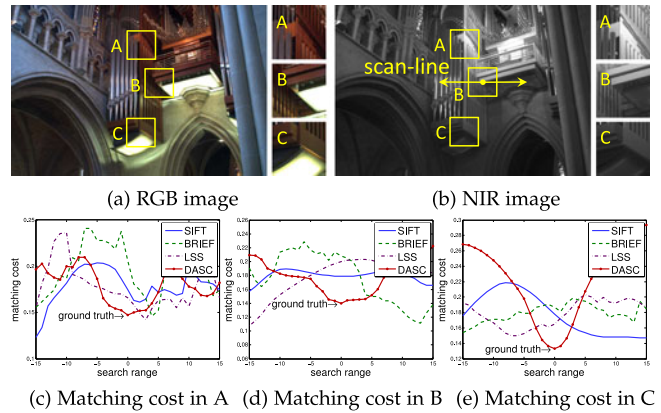(c) Matching cost in A   (d) Matching cost in B   (e) Matching cost in C

Fig. 2. Examples of matching cost comparison. Multi-spectral RGB and NIR images have locally non-linear deformation as depicted in A, B, and C. Matching costs computed with different descriptors along A, B, and C's scan-lines are plotted in (c)-(e). Unlike conventional descriptors, the proposed DASC descriptor yields a reliable global minimum.

encodes an intrinsic structure while providing the robustness against modality variations. To further improve the matching quality and runtime efficiency, we propose a randomized receptive field pooling strategy using sampling patterns that select two patches within the local support window. A linear discriminative learning is employed for obtaining an optimal sampling pattern. The computational redundancy that arises when computing densely sampled descriptors over an entire image is dramatically reduced by applying fast edge-aware filtering [21].

Furthermore, in order to address geometric variation problems, we propose the geometry-invariant DASC (GI-DASC) descriptor that leverages the efficiency and effectiveness of the DASC through a superpixel-based representation. Specifically, we infer an initial geometric field with corresponding scale and rotation of reliable sparse keypoints obtained using weighted maximally self-dissimilarity (WMSD), and then propagate the initial geometric field on a superpixel graph. After transforming sampling patterns according to geometric fields on each superpixel, the DASC is efficiently computed with the transformed sampling patterns on each superpixel extended subimage. Compared to conventional geometry-invariant methods [22], [23], which have been focusing on employing powerful optimization schemes, the GI-DASC provides geometric and photometric robustness on the descriptor itself.

Experimental results show that the DASC outperforms conventional area-based and feature-based approaches on various benchmarks including modality variations; (1) Middlebury stereo benchmark containing illumination and exposure variations [24], (2) multi-modal and multi-spectral dataset including RGB-NIR images [1], [9], different exposure [8], [9], flash-noflash images [8], blurry images [6], [7], and RGB-depth images [9], (3) non-rigid deformation benchmark [25], and (4) MPI optical flow benchmark containing specular reflections, motion blur, and defocus blur [11]. We also show that the GI-DASC outperforms existing geometry-invariant methods on a novel multi-modal benchmark.

1.1 Contribution

The contributions of this paper can be summarized as follows. First, to the best of our knowledge, our approach is the first attempt to design an efficient, dense descriptor for matching multi-modal and multi-spectral images, even under varying geometric conditions. Second, unlike a center-biased dense max pooling, we propose a randomized receptive field pooling with sampling patterns optimized via a discriminative learning, making the descriptor more robust to matching outliers incurred by different imaging modalities. Third, we propose an efficient computational scheme that significantly improves the runtime efficiency of the proposed dense descriptor. Fourth, a geometry-invariant dense descriptor is also proposed, which provides a geometric robustness as a descriptor itself.

This manuscript extends its preliminary version [26]. It newly adds (1) a scale and rotation invariant extension of the DASC, called GI-DASC; (2) a new multi-modal benchmark with a ground truth annotation, captured under varying photometric and geometric conditions; and (3) an intensive comparative study with existing geometry invariant methods using various datasets. The source code of our work (including DASC and GI-DASC) and the new multi-modal benchmark are available at our project webpage [27].

2 RELATED WORK

2.1 Feature Descriptors

As a pioneering work, the scale invariant feature transform (SIFT) was first introduced by Lowe [28] to estimate robust sparse feature correspondence under geometric and photometric variations. Based on the intensity comparison, fast binary descriptors, such as binary robust independent elementary features (BRIEF) [29] and fast retina keypoint (FREAK) [30], have been proposed. Unlike these sparse descriptors, Tola et al. developed a dense descriptor, called DAISY [13], which re-designs conventional sparse descriptors, i.e., SIFT, to efficiently compute densely sampled descriptors over an entire image. Although these conventional gradient-based and intensity comparison-based descriptors show satisfactory performance for small photometric deformation, they cannot properly describe multi-modal and

multi-spectral images that often exhibit severe non-linear deformation.

To estimate correspondences in multi-modal and multi-spectral images, some variants of the SIFT have been developed [31], but these gradient-based descriptors have an inherent limitation similar to the SIFT, especially when an image gradient varies across different modality images. Simo-Serra et al. proposed the deformation and light invariant (DaLI) descriptor [25], but it has a difficulty in providing dense descriptors due to its heavy computational time. Schechtman and Irani introduced the LSS descriptor [18] for the purpose of template matching, and achieved impressive results in object detection and retrieval. Torabi et al. employed the LSS as a multi-spectral similarity metric to register human region of interests (ROIs) [20]. The LSS also has been applied to the registration of multi-spectral remote sensing images [32]. For multi-modal medical image registration, Heinrich et al. proposed a modality independent neighborhood descriptor (MIND) [19] inspired by the LSS. However, none of these approaches scale very well to dense matching tasks for multi-modal and multi-spectral images due to a low discriminative power and a huge complexity.

Recently, several approaches started to employ deep convolutional neural networks (CNNs) [33] for estimating correspondences. For designing explicit, discriminative feature descriptors, intermediate activations from CNN architecture are extracted [34], [35], [36], [37], and they have been shown to be effective for patch-level tasks. However, even though CNN-based descriptors encode a discriminative structure with a deep architecture, they have inherent limitations in multi-modal images, since they use shared convolutional kernels across images which lead to inconsistent responses similar to conventional descriptor [37], [38]. Furthermore, they are unable to provide dense descriptors in the image due to a prohibitively high computational complexity.

2.2 Area-Based Similarity Measures

As surveyed in [39], the mutual information (MI), leveraging the entropy of the joint probability distribution function (PDF), has been popularly applied to a registration of multi-modal medical images. However, the MI is sensitive to local radiometric variation since it formulates the intensity variation in a global manner using the joint entropy computed over an entire image. In [40], this issue can be alleviated to some extent by leveraging a locally adaptive weight obtained from SIFT matching, called MI+SIFT in this paper, but its performance is still limited against the multi-modal variation [41]. Although cross-correlation based methods such as an adaptive normalized cross-correlation (ANCC) [42] show satisfactory results for locally linear variations, they show a limitation under severe modality variations. Irani et al. employed the cross-correlation on the Laplacian energy map for measuring multi-sensor image similarity [43], but it also shows a limitation for general image matching tasks. A robust selective normalized cross-correlation (RSNCC) [9] was proposed for the dense alignment between multi-modal images, but its performance is still unsatisfactory due to an inherent limitation of intensity based similarity measure.

2.3 Geometry-Invariant Dense Correspondences

Based on the SIFT flow (SF) [14] optimization, many methods have been proposed to alleviate geometric variation problems, including deformable spatial pyramid (DSP) [15], scale-less SIFT flow (SLS) [44], scale-space SIFT flow (SSF) [45], and generalized DSP (GDSP) [23]. However, they have a critical limitation as huge computational complexity derived from dramatically large search space in geometry-invariant dense correspondence. A generalized PatchMatch (GPM) [46] was proposed for efficient matching leveraging a randomized search scheme. The DAISY Filter Flow (DFF) [22], which exploits DAISY descriptor [13] with PatchMatch Filter (PMF) [47], was proposed to provide geometric invariance. However, their weak spatial smoothness often induces mismatched results. The scale invariant descriptor (SID) [48] was proposed to encode geometric robustness on the descriptor itself, but it is not tailored to multi-modal matching. Segmentation-aware approach [49] was proposed to provide geometric robustness for descriptors, e.g., SIFT [28] or SID [48], but it may have a negative effect on the discriminative power of the descriptor.

3 BACKGROUND

Let us define an image as $f_i : \mathcal{I} \to \mathbb{R}$ for pixel i, where $\mathcal{I} \subset \mathbb{N}^2$ is a discrete image domain. Given the image f_i, a dense descriptor $\mathcal{D}_i : \mathcal{I} \to \mathbb{R}^L$ is defined on a local support window $\mathcal{R}_i$ centered at pixel i with a feature dimension L. Conventionally, many descriptors were computed based on the assumption that there is a common underlying visual pattern which is shared by two images. However, as shown in Fig. 2, multi-spectral images such as a pair of RGB-NIR have a nonlinear photometric deformation even within a small window, e.g., gradient reverse and intensity order variation. More seriously, there are outliers including structure divergence caused by shadow or highlight. In these cases, conventional descriptors using an image gradient (SIFT [28]) or an intensity comparison (BRIEF [29]) cannot capture coherent matching evidences, resulting erroneous local minima in estimating dense correspondences.

Unlike these conventional descriptors, the LSS descriptor $\mathcal{D}_i^{\text{lss}}$ measures a correlation between two patches $\mathcal{F}_i$ and $\mathcal{F}_j$ centered at two pixels i and j within a local support window $\mathcal{R}_i$ [18]. As shown in Fig. 3a, it discretizes the correlation surface on a log-polar grid, generates a set of bins, and then stores a maximum correlation value within each bin. Formally, $\mathcal{D}_i^{\text{lss}} = \bigcup_l d_{i,l}^{\text{lss}}$ for $l = 1, \ldots, L^{\text{lss}}$ is a $L^{\text{lss}} \times 1$ feature vector, and $d_{i,l}^{\text{lss}}$ can be computed as follows:

$$d_{i,l}^{\text{lss}} = \max_{j \in \text{bin}_i(l)} \{ \mathcal{C}(i,j) \}, \tag{1}$$

where $\text{bin}_i(l) = \{ j | j \in \mathcal{R}_i, \rho_{r-1} < |i - j| \leq \rho_r, \theta_{a-1} < \angle(i - j) \leq \theta_a \}$ with a log radius ρ_r for $r \in \{1, \ldots, N_\rho\}$ and a quantized angle θ_a for $a \in \{1, \ldots, N_\theta\}$ with $\rho_0 = 0$ and $\theta_0 = 0$. In that case, $L^{\text{lss}} = N_\rho \times N_\theta$. The correlation surface $\mathcal{C}(i,j)$ is typically computed using a simple similarity metric such as the sum of squared difference (SSD) with a normalization factor σ_s

$$\mathcal{C}(i,j) = \exp(-\text{SSD}(\mathcal{F}_i, \mathcal{F}_j)/\sigma_s). \tag{2}$$

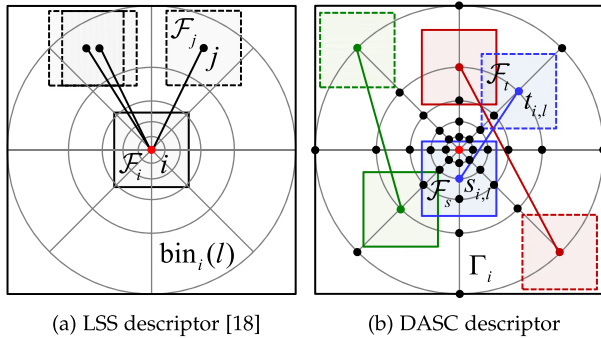(a) LSS descriptor [18]   (b) DASC descriptor

Fig. 3. Demonstration of the LSS [18] and the DASC descriptor. Within the support window, solid and dotted line box depict source and target patch, respectively. Unlike a center-biased dense max pooling on each $\mathrm{bin}_i(l)$ in the LSS descriptor, the DASC descriptor incorporates a randomized receptive field pooling using sampling pattern $(s_{i,l}, t_{i,l}) \in \Lambda_i^{\mathrm{dasc}}$ on Γ_i, optimized by a discriminative learning.

This LSS descriptor has been shown to be robust in cross-domain object detection [18], but it provides unsatisfactory results in densely matching multi-modal images as shown in Fig. 2. It is because the max pooling strategy performed in each $\mathrm{bin}_i(l)$ loses matching details, leading to a poor discriminative power. Furthermore, the center-biased correlation measure cannot handle severe outliers effectively, which frequently exist in multi-modal and multi-spectral images. In terms of a computational complexity, there exists no efficient computational scheme designed for dense matching descriptor.

4 THE DASC DESCRIPTOR

4.1 Randomized Receptive Field Pooling

Instead of using a center-biased max pooling of the LSS descriptor in Fig. 3a, our DASC descriptor incorporates a randomized receptive field pooling with sampling patterns in such a way that a pair of two patches are randomly selected within a local support window. It is motivated by three observations; 1) In multi-spectral and multi-modal images, there frequently exist non-informative regions which are locally degraded, e.g., shadows or outliers. 2) Center-biased pooling is very sensitive to a degradation of a center patch, and cannot deal with a homogeneous or salient center pixel which does not contain self-similarities [18]. 3) From the relationship between Census transform [50] and BRIEF [29] descriptor, it is shown that the randomness enables a descriptor to encode structural information more robustly.

Our approach encodes a similarity between patch-wise receptive fields sampled from log-polar circular point set Γ_i as shown in Fig. 3b. It is defined as $\Gamma_i = \{j | j \in \mathcal{R}_i, |i - j| = \rho_r, \angle(i - j) = \theta_a\}$ where the number of points is defined as $N_c = N_\rho \times N_\theta + 1$, and has a higher density of points near a center pixel, similar to DAISY descriptor [13]. Given N_c points in Γ_i, there exist $N_{pc} = \{N_c \times (N_c - 1)\}/2$ candidate sampling patterns, leading to a dramatically high-dimension descriptor. However, many of the sampling pattern pairs might not be useful in describing a local support window. Therefore, we employ a randomized approach to extract L^{dasc} sampling patterns from N_{pc} pattern candidates. Our descriptor $\mathcal{D}_i^{\mathrm{dasc}} = \bigcup_l d_{i,l}^{\mathrm{dasc}}$ for $l = 1, \ldots, L^{\mathrm{dasc}}$ is encoded

with a set of patch similarity between two patches based on sampling patterns that are selected from Γ_i

$$d_{i,l}^{\mathrm{dasc}} = \mathcal{C}(s_{i,l}, t_{i,l}), \quad s_{i,l}, t_{i,l} \in \Gamma_i, \tag{3}$$

where $s_{i,l}$ and $t_{i,l}$ are lth selected sampling patterns at pixel i. Note that the sampling patterns are fixed for all pixels in an image. Namely, all pixels share the same set of offset vectors $t_{i,l} - s_{i,l}$ for $l = 1, \ldots, L^{\mathrm{dasc}}$, enabling a fast computation of dense descriptors, which will be detailed in Section 4.3. Although the DASC descriptor uses only sparse patch-wise pairs in a local support window, many of patches are overlapped when computing patch similarities between the sparse pairs, allowing the descriptor to consider the majority of pixels in the support window and reflect original image attributes effectively.

4.1.1 Sampling Pattern Learning

Finding an optimal sampling pattern is a critical issue in the DASC descriptor. With the assumption that there is no single hand-craft feature that always provides the robustness to all circumstances [51], we employ a discriminative learning to obtain optimal sampling patterns within a local support window. Given candidate sampling patterns $\Lambda_i = \{(s_{i,l}, t_{i,l}) | l = 1, \ldots, N_{pc}\}$, our goal is to select the best sampling patterns which derive an important spatial layout.

Our approach exploits support vector machines (SVMs) with a linear kernel [52]. For learning, we build a dataset $\mathcal{P} = \{(\mathcal{R}_h^1, \mathcal{R}_h^2, y_h) | h = 1, \ldots, N_{tr}\}$, where $(\mathcal{R}^1, \mathcal{R}^2)$ are support window pairs in multi-modal or multi-spectral images, and N_{tr} is the number of training samples. y is a binary label that becomes 1 if two patches are matched, or 0 otherwise. The training data set $\mathcal{P}$ was built with images captured under varying illumination conditions and/or with imaging devices [9], [11], [24]. In experiments, $N_{tr} = 10{,}000$.

First, the feature $\mathbf{r}_h = \bigcup_l r_{h,l}$ that describes two support window pairs $\mathcal{R}_h^1$ and $\mathcal{R}_h^2$ is defined

$$r_{h,l} = \exp\left(-\left(d_{h,l}^{\mathrm{dasc},1} - d_{h,l}^{\mathrm{dasc},2}\right)^2 / 2\sigma_r^2\right), \tag{4}$$

where σ_r is a Gaussian parameter, and $d_{h,l}^{\mathrm{dasc}}$ is the DASC descriptor. The decision function $\mathcal{Q}$ to classify training dataset $\mathcal{P}$ into matching and non-matching can be represented as

$$\mathcal{Q}(\mathbf{r}_h) = \mathbf{v}^T \mathbf{r}_h + \mathbf{b}, \tag{5}$$

where the weight $\mathbf{v} = \bigcup_l v_l$ indicates an amount of contribution of each candidate sampling pattern, and $\mathbf{b}$ is a bias. Learning $\mathbf{v}$ can be formulated as minimizing

$$\mathbf{E}_{\mathrm{svm}}(\mathbf{v}) = \|\mathbf{v}\|^2 + C_{\mathrm{svm}} \sum_{h=1}^{N_{tr}} l_{\mathrm{hinge}}(y_h \cdot \mathcal{Q}(\mathbf{r}_h)), \tag{6}$$

where the hinge loss function $l_{\mathrm{hinge}}(x) = \max(0, 1 - x)$ and C_{svm} represents a regularization parameter. We use LIBSVM [52] to minimize this objective function. The $|v_l|$ encodes the importance of corresponding sampling pattern towards the final decision [53]. Therefore, we rank top L^{dasc} sampling patterns based on $|v_l|$ value, and use them in our descriptor, which is denoted as $\Lambda_i^{\mathrm{dasc}}$.

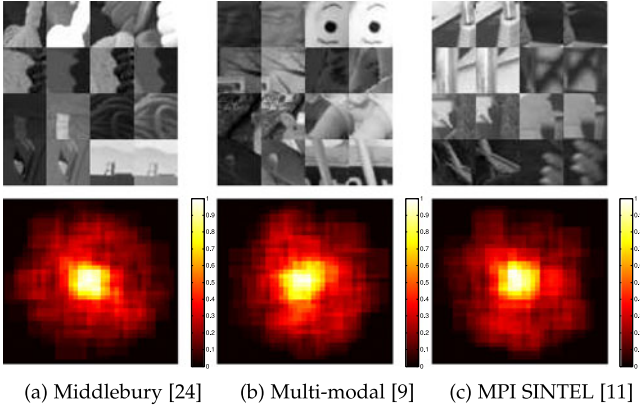(a) Middlebury [24] (b) Multi-modal [9] (c) MPI SINTEL [11]

Fig. 4. Visualization of patch-wise receptive fields of the DASC descriptor learned from the training set $\mathcal{P}$ built with the Middlebury benchmark [24], multi-modal benchmark [9], and the MPI SINTEL benchmark [11]. Similar to [51], we stacked all patch-wise receptive fields learned from each training image, and normalized them with the maximal value.

Fig. 4 visualizes learned patch-wise receptive fields of the DASC. It looks similar to the Gaussian weighting, which has been proven to be effective in terms of a structural encoding of descriptor in many literatures [51], [54]. According to training set, it learns optimal receptive fields.

4.2 Adaptive Self-Correlation Measure

With estimated sampling patterns $(s_{i,l}, t_{i,l})$, the DASC descriptor measures a patch similarity using an adaptive self-correlation measure in order to robustly encode a local internal layout of self-similarities. For the sake of simplicity, we omit (i, l) in the correlation metric from here on, as it is repeatedly computed for all (i, l). For $(s, t) \in \Lambda^{\mathrm{dasc}}$, the adaptive self-correlation $\Psi(s, t)$ between two patches $\mathcal{F}_s$ and $\mathcal{F}_t$ centered at pixels s and t is computed as follows:

$$\Psi(s, t) = \frac{\sum_{s',t'} \omega_{s,s'} \omega_{t,t'} (f_{s'} - \mathcal{G}_s)(f_{t'} - \mathcal{G}_t)}{\sqrt{\sum_{s'} \{\omega_{s,s'}(f_{s'} - \mathcal{G}_s)\}^2} \sqrt{\sum_{t'} \{\omega_{t,t'}(f_{t'} - \mathcal{G}_t)\}^2}}, \quad (7)$$

where $s' \in \mathcal{F}_s$ and $t' \in \mathcal{F}_t$ and weighted averages on $\mathcal{F}_s$ and $\mathcal{F}_t$ are defined as $\mathcal{G}_s = \sum_{s'} \omega_{s,s'} f_{s'}$ and $\mathcal{G}_t = \sum_{t'} \omega_{t,t'} f_{t'}$.

The weight $\omega_{s,s'}$ represents how similar two pixels s and s' are, and is normalized, i.e., $\sum_{s'} \omega_{s,s'} = 1$. It can be defined with any kind of edge-aware weights [21], [55], [56]. This weighted sum better handles outliers and local variations in patches compared to other patch-wise similarity metrics. It is worth noting that the adaptive self-correlation used here is conceptually similar to the ANCC [42], but our descriptor employs the correlation metric for measuring self-similarity within a single image which is used for matching two or more images later, while the ANCC is used to directly measure inter-similarity between different images.

Finally, our patch-wise similarity between $\mathcal{F}_s$ and $\mathcal{F}_t$ is computed with a truncated exponential function, which has been widely used in the robust estimator [57]

$$\mathcal{C}(s, t) = \max(\exp(-(1 - |\Psi(s, t)|)/\sigma_c), \tau_c), \quad (8)$$

where σ_c is a bandwidth of Gaussian kernel and τ_c is a truncation parameter. Here, a absolute value of $\Psi(s, t)$ is used to mitigate the effect of intensity reverses. The correlation $\mathcal{C}(s_{i,l}, t_{i,l})$ for i is normalized with an unit norm for all l.
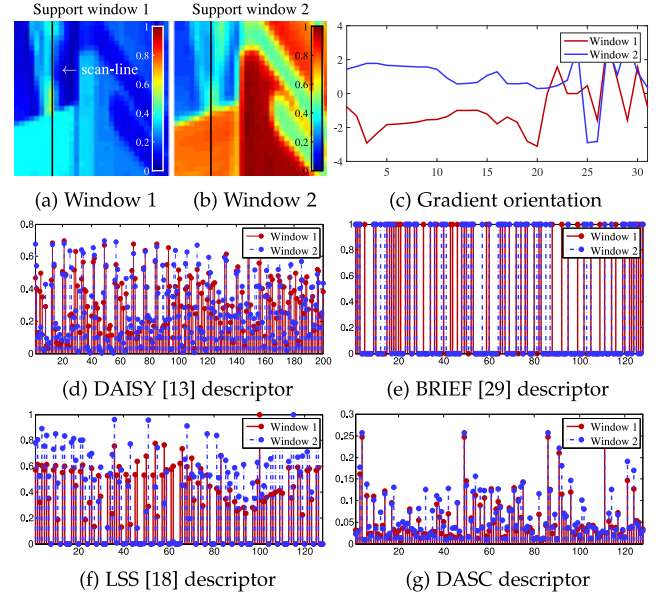

(a) Window 1 (b) Window 2 (c) Gradient orientation

(d) DAISY [13] descriptor (e) BRIEF [29] descriptor

(f) LSS [18] descriptor (g) DASC descriptor

Fig. 5. Visualization of support window pairs on multi-spectral RGB and NIR images denoted as 'A' in Fig. 2 having gradient orientation variations, and descriptors for these window pairs. Conventional descriptors such as DAISY [13], BRIEF [29], and LSS [18] vary across modality variations. Unlike those methods, our DASC descriptor remains unchanged to modality variations.

Fig. 5 represents examples of visualizing the results of various descriptors. The conventional descriptors show the sensitivity to modality variations, however the DASC shows the robustness against multi-modal variations.

4.3 Efficient Computation for Dense Descriptor

For densely constructing our descriptor on an entire image, we should compute $\mathcal{C}(s_{i,l}, t_{i,l})$ for all patch pairs belonging to $(s_{i,l}, t_{i,l}) \in \Lambda_i^{\mathrm{dasc}}$ for each pixel i. Thus, a straightforward computation can be extremely time-consuming. In this section, we present an efficient method for computing the DASC descriptor. To compute all weighted sums in (7) for $(s_{i,l}, t_{i,l})$ efficiently, we employ a constant-time edge-aware filter (EAF), e.g., the guided filter (GF) [21]. However, the symmetric weight $w_{s,s'} w_{t,t'}$ varies for each l, and thus computing the numerator in (7) is still very time-consuming.

To alleviate these limitations, we simplify (7) by considering only the weight $w_{s,s'}$ from the source patch $\mathcal{F}_s$ so that a fast computation of (7) using fast edge-aware filter is feasible. It should be noted that such an asymmetric weight approximation also has been used in cost aggregation for stereo matching [58]. We also found that in our descriptor, a performance gap between using the asymmetric weight $w_{s,s'}$ and the symmetric weight $w_{s,s'} w_{t,t'}$ is negligible, which will be shown in Section 6.2.5. For efficient description, we also re-arrange the sampling pattern $(s_{i,l}, t_{i,l})$ to referenced-biased pairs $(i, j) = (i, i + t_{i,l} - s_{i,l})$. Eq. (7) is then approximated as follows:

$$\tilde{\Psi}(i, j) = \frac{\sum_{i',j'} \omega_{i,i'} (f_{i'} - \mathcal{G}_i)(f_{j'} - \mathcal{G}_{i,j})}{\sqrt{\sum_{i'} \omega_{i,i'} (f_{i'} - \mathcal{G}_i)^2} \sqrt{\sum_{i',j'} \omega_{i,i'} (f_{j'} - \mathcal{G}_{i,j})^2}}, \quad (9)$$

where $\mathcal{G}_i = \sum_{i'} \omega_{i,i'} f_{i'}$. Furthermore, $\mathcal{G}_{i,j} = \sum_{i',j'} \omega_{i,i'} f_{j'}$ which means weighted average of $f_{j'} \in \mathcal{F}_j$ with a guidance image
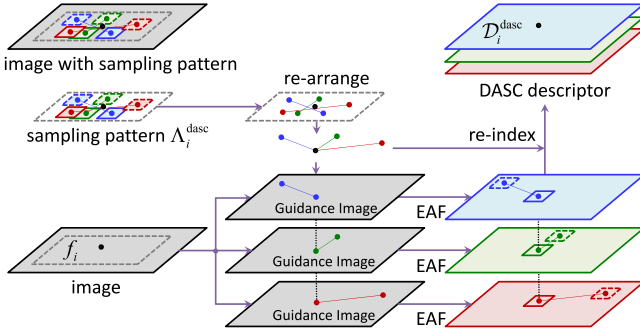
Fig. 6. Efficient computation framework of the DASC descriptor. In order to reduce a computational load in computing the adaptive self-correlation, it re-arranges the sampling pattern and employs fast EAF scheme. The DASC descriptor is then computed with re-indexing.

$f_{i'} \in \mathcal{F}_i$. It is worth noting that the robustness of $\Psi(s,t)$ can be still applied to $\tilde{\Psi}(i,j)$ since their difference is just weight factors.

We then decompose numerator and denominator in (9) after some arithmetic derivations such that

$$\frac{\mathcal{G}_{i,ij} - \mathcal{G}_i \cdot \mathcal{G}_{i,j}}{\sqrt{\mathcal{G}_{i^2} - \mathcal{G}_i^2} \cdot \sqrt{\mathcal{G}_{i,j^2} - \mathcal{G}_{i,j}^2}}, \qquad (10)$$

where $\mathcal{G}_{i^2} = \sum_{i'} \omega_{i,i'} f_{i'}^2$, $\mathcal{G}_{i,ij} = \sum_{i',j'} \omega_{i,i'} f_{i'} f_{j'}$, and $\mathcal{G}_{i,j^2} = \sum_{i',j'} \omega_{i,i'} f_{j'}^2$. While the $\mathcal{G}_i$ and $\mathcal{G}_{i^2}$ can be computed on image domain once, $\mathcal{G}_{i,ij}$, $\mathcal{G}_{i,j}$, and $\mathcal{G}_{i,j^2}$ should be computed on each offset. However, the weight $\omega_{i,i'}$ is fixed for all offsets, thus it can be shared in all offsets. All these components can be efficiently computed using a constant-time edge-aware filter [21]. Finally, the dense descriptor $\mathcal{D}_i^{\text{dasc}}$ is computed with re-indexing as $d_{i,l}^{\text{dasc}} = \mathcal{C}(s_{i,l}, t_{i,l})$ though the robust function in (8). Fig. 6 describes our efficient method for computing the DASC descriptor. Algorithm 1 summarizes the efficient computation of the DASC descriptor.

Algorithm 1. Dense Adaptive Self-Correlation (DASC)

Input: image f_i, candidate sampling patterns Λ_i, training patch pairs dataset $\mathcal{P}$.
Output: the DASC descriptor volume $\mathcal{D}_i^{\text{dasc}}$.
  /∗ *Offline Procedure* ∗/
1:   Compute $\mathbf{r}_h$ using (4) for possible candidate sampling patterns Λ_i on training support window pairs $\mathcal{P}$.
2:   Learn a weight $\mathbf{v}$ by optimizing (6).
3:   Select the maximal L^{dasc} sampling patterns $(s_{i,l}, t_{i,l})$ in terms of $|v_l|$, denoted as Λ_i^{dasc}. dasc.
  /∗ *Online Procedure* ∗/
4:   Compute $\mathcal{G}_i = \sum_{i'} \omega_{i,i'} f_{i'}$ for all pixel i.
5:   Compute $\mathcal{G}_{i^2} = \sum_{i'} \omega_{i,i'} f_{i'}^2$.
    for $l = 1 : L^{\text{dasc}}$ **do**
6:    Re-arrange $(s_{i,l}, t_{i,l}) \in \Lambda_i^{\text{dasc}}$ as $(i,j) = (i, i + t_{i,l} - s_{i,l})$.
7:    Compute $\mathcal{G}_{i,ij} = \sum_{i',j'} \omega_{i,i'} f_{i'} f_{j'}$.
8:    Compute $\mathcal{G}_{i,j} = \sum_{i',j'} \omega_{i,i'} f_{j'}$.
9:    Compute $\mathcal{G}_{i,j^2} = \sum_{i',j'} \omega_{i,i'} f_{j'}^2$.
10:   Estimate $\tilde{\Psi}(i,i')$ and $\mathcal{C}(i,i')$ using (9) and (8).
11:   Compute the DASC descriptor $\mathcal{D}_i^{\text{dasc}}$ by re-indexing sampling patterns such that $d_{i,l}^{\text{dasc}} = \mathcal{C}(s_{i,l}, t_{i,l})$.
    end for

4.3.1 Comparison of Symmetric and Asymmetric Version of Adaptive Self-Correlation Measure

This section analyzes the performance of the DASC descriptor when using the symmetric weight $\omega_{s,s'} \omega_{t,t'}$ of $\Psi(s,t)$ in (7) and with the asymmetric weight $\omega_{i,i'}$ of $\tilde{\Psi}(i,j)$ in (9). The symmetric weight case in the DASC can also be computed similar to Section 4.3. After re-arranging the sampling pattern as $(i,j) = (i, i + t_{i,l} - s_{i,l})$, the (7) can be then decomposed as similar in (10)

$$\Psi(i,j) = \frac{\mathcal{G}_{ij,ij} - \mathcal{G}_{ij,i} - \mathcal{G}_{ij,j} + \mathcal{G}_i \cdot \mathcal{G}_j}{\sqrt{\mathcal{G}_{i^2} - \mathcal{G}_i^2} \cdot \sqrt{\mathcal{G}_{j^2} - \mathcal{G}_j^2}}, \qquad (11)$$

where $\mathcal{G}_{ij,ij} = \sum_{i'} \omega_{i,i'} \omega_{j,j'} f_{i'}^2 f_{j'}^2$, $\mathcal{G}_{ij,i} = \sum_{i',j'} \omega_{i,i'} \omega_{j,j'} f_{i'}$, and $\mathcal{G}_{ij,j} = \sum_{i',j'} \omega_{i,i'} \omega_{j,j'} f_{j'}$. The denominator can be easily computed on overall image once. However, compared to the asymmetric measure in (9), $\omega_{i,i'} \omega_{j,j'}$ in $\mathcal{G}_{ij,ij}$, $\mathcal{G}_{ij,i}$, and $\mathcal{G}_{ij,j}$ varies for each l. Furthermore, it should be computed with a range distance using 6-D vector (or 2-D vector), when an input is a color image (or an intensity image). It significantly increases a computational burden needed for employing constant-time EAFs [21], [59]. A performance gap between using the symmetric measure $\Psi(s,t)$ and the asymmetric measure $\tilde{\Psi}(i,j)$ in the DASC descriptor is negligible, which will be shown in Section 6.2.5.

4.4 Computational Complexity Analysis

The computational complexity of the DASC descriptor on the brute-force implementation becomes $O(INL)$, where I, N, and L represent an image size, a patch size, and a descriptor dimension, respectively. With our efficient computation model, our approach removes the complexity dependency on the patch size N, i.e., $O(IL)$ due to fast constant-time EAF. Furthermore, since there exist repeated offsets, the complexity is further reduced as $O(I\tilde{L})$ for $\tilde{L} < L$.

5 GEOMETRY-INVARIANT DASC DESCRIPTOR

Similar to the DAISY [13], the DASC descriptor is not appropriate to deal with geometric variations. In this section, we propose the geometry-invariant DASC descriptor, called GI-DASC, that addresses severe geometric variations as well as image modality variations. A key idea is to geometrically transform sampling patterns used to measure the patch similarity according to scale and rotation fields when computing the DASC descriptor. To estimate the scale and rotation fields, we first infer initial geometric fields only for sparse points. These initial fields are then fitted and propagated through a superpixel graph. Finally, the GI-DASC descriptor is efficiently computed with geometrically transformed sampling patterns in a manner similar to computing the DASC descriptor, except the fact that the descriptor computation is done for each superpixel independently.

Adopting the superpixel-based geometry field inference has the following three reasons. First, the reliable geometry field can be estimated reliably only at distinctive pixels. Second, the geometric fields tend to vary smoothly, except object boundaries. Third, the transformed sampling patterns should be fixed for each superpixel so that the computational scheme based on the fast EAF [21] can be used for
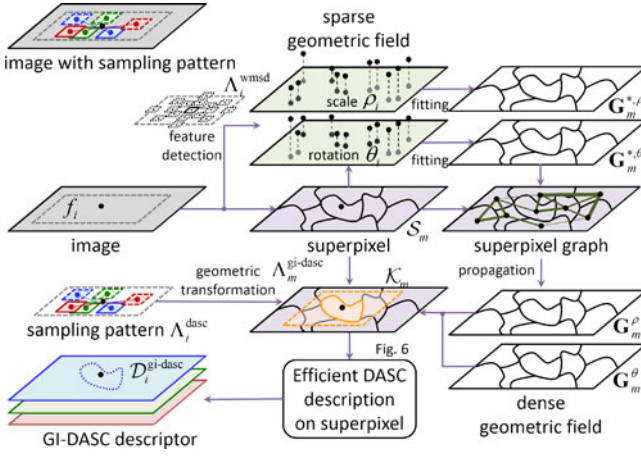
Fig. 7. Efficient computation framework of the geometry-invariant DASC (GI-DASC) descriptor. To leverage the efficient computation scheme of the DASC, we employ a superpixel-based description with inferred geometric fields on each superpixel using the WMSD detection.

efficiently obtaining the GI-DASC for each superpixel. Fig. 7 represents the overview of the GI-DASC.

5.1 Initial Sparse Geometric Field Inference

Conventional feature detectors, e.g., SIFT [28], are very sensitive to multi-modal and multi-spectral deformation. In order to extract sparse features with distinctive geometric information available, we employ maximal self-dissimilarity (MSD) thanks to its robustness for modality deformation [60]. We propose weighted MSD (WMSD) that improves the performance of the MSD in terms of both complexity and robustness by employing an weighted similarity measure and an efficient computation scheme similar to the DASC.

Similar to Γ_i used in the DASC, the log-polar circular point set Γ_i^{wmsd} is defined for feature detector. The sampling pattern $\Lambda_i^{\mathrm{wmsd}}$ is then defined in such a way that the source patch is always located at center pixel and the target patches are located at other neighboring points as shown in Fig. 8a. In order to consider the scale deformation, we build the Gaussian image pyramid $u_i^k = f_i \times \varrho_k$ for $k = 1, \ldots, N_k$, where ϱ_k is the kth Gaussian kernel with a sigma ρ_k and N_k is the number of pyramids. After re-arranging the sampling pattern as $(i, j) = (i, i + t_{i,l} - s_{i,l})$, The self-dissimilarity measure $\Phi^k(i, l)$ for $l = 1, \ldots, L^{\mathrm{wmsd}}(= N_\rho^{\mathrm{wmsd}} \times N_\theta^{\mathrm{wmsd}})$ is computed using weighted sum of squared difference (SSD) with a guidance image $u_{i'}^k$ such that

$$\Phi^k(i, l) = \sum_{i',j'} \omega_{i,i'} (u_{i'}^k - u_{j'}^k)^2 \qquad (12)$$
$$= \mathcal{U}_{i2}^k + \mathcal{U}_{i,j2}^k - 2\mathcal{U}_{i,ij}^k,$$

where $\mathcal{U}_{i2}^k = \sum_{i'} \omega_{i,i'} (u_{i'}^k)^2$, $\mathcal{U}_{i,j2}^k = \sum_{i',j'} \omega_{i,i'} (u_{j'}^k)^2$, and $\mathcal{U}_{i,ij}^k = \sum_{i',j'} \omega_{i,i'} u_{i'}^k u_{j'}^k$. Similar to the DASC, (12) can be computed efficiently using constant time EAF [21], [59].

We extract the index set Π_i^o for the o most smallest value $\Psi_{i,l}^{\mathrm{wmsd},k}$ for all l, i.e., o nearest neighbors for center patch in Fig. 8b. It should be noted that parameter o trades distinctiveness and computational efficiency [60]. We then compute feature response map Ω_i^k by estimating the summation of $\Phi^k(i, l)$ for $l \in \Pi_i^o$ such that
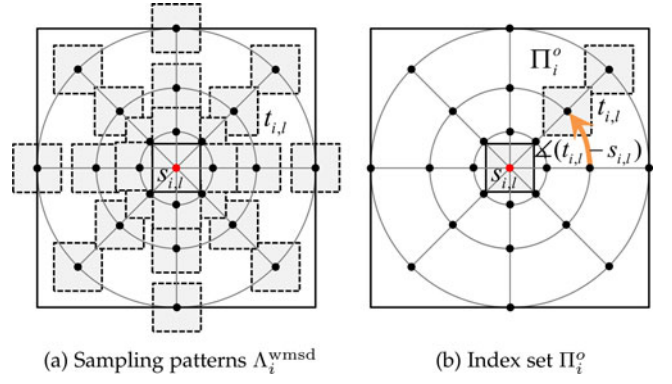

(a) Sampling patterns $\Lambda_i^{\mathrm{wmsd}}$ (b) Index set Π_i^o

Fig. 8. Demonstration of sampling patterns $(s_{i,l}, t_{t,l}) \in \Lambda_i^{\mathrm{wmsd}}$ for the WMSD detector and the index set for the o most smallest value Π_i^o. It enables us to extract reliable feature points $i \in \mathcal{I}'$ with corresponding geometric fields (scale ρ_i and rotation θ_i).

$$\Omega_i^k = \sum_{l \in \Pi_i^o} \Phi^k(i, l). \qquad (13)$$

For feature response maps $\mathbf{\Omega}_i = \{\Omega_i^k\}$, the local maxima are obtained by the non maximal suppression, which compares Ω_i^k to its 8 neighbors on the current scale and 18 neighbors on the $(k + 1)$th and $(k - 1)$th scales. Similar to SIFT [28], a feature point $i \in \mathcal{I}'$ is detected only if $\{\Omega_i^k\}$ has an extreme value compared to all of these neighbors, and its scale ρ_i is defined with ρ_k, where $\mathcal{I}' \subset \mathcal{I}$ is a sparse discrete image domain.

A canonical orientation is further associated to $i \in \mathcal{I}'$ by constructing a histogram with angles $\measuredangle(t_{i,l} - s_{i,l})$ for $l \in \Pi_i^o$ weighted by $\Phi^k(i, l)$ as

$$l_{\mathrm{hist}}(i, \theta) = \sum_{l \in \Pi_i^o} \Phi^k(i, l) \cdot \delta(\measuredangle(t_{i,l} - s_{i,l}) - \theta), \qquad (14)$$

where δ is the Kronecker delta function. Then, we simply choose the direction corresponding to the highest bin in the histogram, i.e., $\theta_i = \arg\max_\theta l_{\mathrm{hist}}(i, \theta)$. The WMSD detector is summarized in Algorithm 2.

Algorithm 2. Weighted Maximal Self-Dissimilarity (WMSD)

Input: image f_i, feature detection sampling patterns Λ_i^{det}.
Output: feature points $i \in \mathcal{I}'$ with scale ρ_i, rotation θ_i.
    for $k = 1 : N_k$ **do**
1:     Compute $u_i^k = f_i * \varrho_k$ with the Gaussian kernel ϱ_k.
2:     Compute $\mathcal{U}_{i2}^k = \sum_{i'} \omega_{i,i'} (u_{i'}^k)^2$ for all pixel i.
       for $l = 1 : L_{\mathrm{wmsd}}$ **do**
3:       Compute $\mathcal{U}_{i,j2}^k = \sum_{i',j'} \omega_{i,i'} (u_{j'}^k)^2$ for $j = i + t_{i,l} - s_{i,l}$.
4:       Compute $\mathcal{U}_{i,ij}^k = \sum_{i',j'} \omega_{i,i'} u_{i'}^k u_{j'}^k$.
5:       Estimate $\Phi^k(i, l) = \mathcal{U}_{i2}^k + \mathcal{U}_{i,j2}^k - 2\mathcal{U}_{i,ij}^k$.
       end for
6:     Extract the index set Π_i^o among $\Phi^k(i, l)$ for all l.
7:     Build response map as $\Omega_i^k = \sum_{l \in \Pi_i^o} \Phi^k(i, l)$.
    end for
8: Detect feature points $i \in \mathcal{I}'$ from $\mathbf{\Omega} = \{\Omega^k\}$ with scale factor ρ_i.
9: Compute the orientation θ_i for i from $l_{\mathrm{hist}}(i, \theta)$.

(a) Image 1   (b) Image 2   (c) Superpixel 1   (d) Superpixel 2

(e) $\mathbf{G}_1^{*,\rho}$ (f) $\mathbf{G}_2^{*,\rho}$ (g) $\mathbf{G}_1^{*,\theta}$ (h) $\mathbf{G}_2^{*,\theta}$

(i) $\mathbf{G}_1^{\rho}$ (j) $\mathbf{G}_2^{\rho}$ (k) $\mathbf{G}_1^{\theta}$ (l) $\mathbf{G}_2^{\theta}$
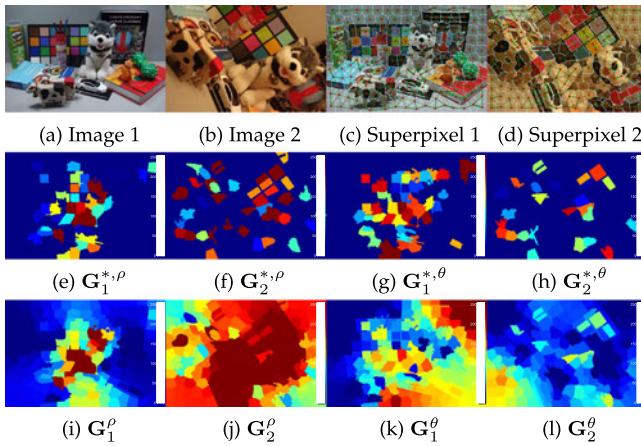
Fig. 9. Examples of a superpixel graph-based propagation. With each superpixel graph in (c), (d) for input images in (a), (b), sparse geometric fields (scale $\mathbf{G}^{*,\rho}$, rotation $\mathbf{G}^{*,\theta}$) in (e)-(h) are propagated into dense geometric fields (scale $\mathbf{G}^{\rho}$, rotation $\mathbf{G}^{\theta}$) in (i)-(l).

5.2 Superpixel Graph-Based Propagation

In order to infer dense geometric fields from sparse geometric fields (ρ_i and θ_i for $i \in \mathcal{I}'$), we decompose the image f as superpixel $\mathcal{S} = \{\mathcal{S}_m | \bigcup_m \mathcal{S}_m = \mathcal{I}$ and $\forall m \neq n, \mathcal{S}_m \bigcap \mathcal{S}_n \neq \varnothing$, $m \in 1, \ldots, N_m\}$, where N_m is the number of superpixels. The geometric field $\mathbf{G}_m^{*,\rho}$ and $\mathbf{G}_m^{*,\theta}$ are fitted on each superpixel $\mathcal{S}_m$ as the average of sparse geometric fields ρ_i and θ_i for $i \in \{\mathcal{I}' \bigcap \mathcal{S}_m\}$. Note that this fitting operation is performed only when $\{\mathcal{I}' \bigcap \mathcal{S}_m\}$ exists, i.e., the superpixel includes sparse feature points (at least, 1). Finally, the $\mathbf{G}^{*,\rho} = \bigcup_m \mathbf{G}_m^{*,\rho} \in \mathbb{R}^{N_m}$ and $\mathbf{G}^{*,\theta} = \bigcup_m \mathbf{G}_m^{*,\theta}$ are constructed for all superpixels.

Similar to [61], our approach then formulates an inference of dense geometric fields $\mathbf{G}^{\rho}$ and $\mathbf{G}^{\theta}$ as a constrained optimization problem where surface-fitted sparse geometric fields $\mathbf{G}^{*,\rho}$ and $\mathbf{G}^{*,\theta}$ are interpreted as soft constraints. For the sake of simplicity, we omit ρ and θ since they can be computed using the same method. The energy function of our superpixel-based propagation is defined as follows:

$$\sum_m \left\{ p_m^{\text{sp}}(\mathbf{G}_m - \mathbf{G}_m^*)^2 + \mu \sum_{n \in \mathcal{N}_m} \omega_{mn}^{\text{sp}}(\mathbf{G}_m - \mathbf{G}_n)^2 \right\}, \qquad (15)$$

where μ is a regularization parameter. Here, the first term encodes the dissimilarity between final geometric fields $\mathbf{G}_m$ and initial sparse geometric fields $\mathbf{G}_m^*$. p_m^{sp} is an index function, which is 1 for valid (constraint) superpixel, and 0 otherwise. The second term imposes the constraint that two adjacent superpixels m and $n \in \mathcal{N}_m$ may have similar geometric fields according to surperpixel feature affinity ω_{mn}^{sp}, which will be described in the following section.

5.2.1 Superpixel Feature Affinity

Our approach employs a superpixel feature composed of an appearance and a spatial feature. First, appearance feature υ_m^c is defined as the average and standard deviation for intensities of pixels within superpixels. In experiments, we used RGB, Lab, and YCbCr space for a color image, thus $\upsilon_m^c \in \mathbb{R}^{18}$. For an NIR image, appearance feature is defined

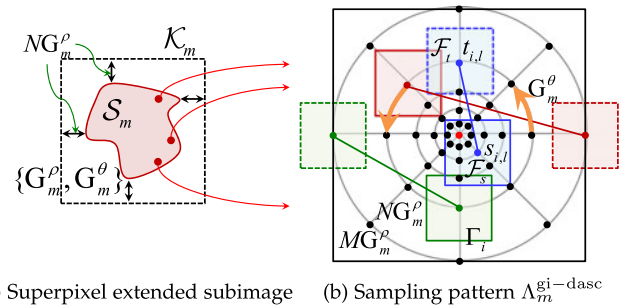(a) Superpixel extended subimage   (b) Sampling pattern $\Lambda_m^{\text{gi-dasc}}$

Fig. 10. Sampling pattern transformation in the GI-DASC descriptor. The sampling patterns $(s_{i,l}, t_{i,l}) \in \Lambda_i^{\text{dasc}}$ is transformed as $(s_{m,l}, t_{m,l}) \in \Lambda_m^{\text{gi-dasc}}$ with $\mathbf{G}_m^{\rho}$ and $\mathbf{G}_m^{\theta}$ on superpixel $\mathcal{S}_m$, which is applied equally for all $i \in \mathcal{S}_m$. It provides the geometric robustness on each superpixel.

on one-channel intensity domain such that $\upsilon_m^c \in \mathbb{R}^2$. Note that directly constructing an affinity matrix with intensity values may lead to inaccurate results due to intensity variations. However, the effect on such variations can be greatly reduced, since the appearance feature is defined as an aggregated form within a superpixel and the affinity value is measured within the same image domain. Second, spatial feature $\upsilon_m^p \in \mathbb{R}^2$ is defined as a spatial centroid coordinate within superpixels. Based on these superpixel features, a superpixel feature affinity ω_{mn}^{sp} between two adjacent superpixel m and $n \in \mathcal{N}_m$ is computed as

$$\omega_{mn}^{\text{sp}} = \exp(-\|\upsilon_m^c - \upsilon_n^c\|^2/\lambda_c - \|\upsilon_m^p - \upsilon_n^p\|^2/\lambda_p), \qquad (16)$$

where λ_c and λ_p denote coefficients for controlling the spatial coherence of neighboring superpixels.

5.2.2 Solver

The minimum of the energy function (15) can be obtained with the following linear system

$$(\mathbf{P} + \mu\mathbf{U} - \mu\mathbf{W})\mathbf{G} = \mathbf{P}\mathbf{G}^*, \qquad (17)$$

where $\mathbf{P}_{mm} = \text{diag}[p_1^{\text{sp}}, \ldots, p_{N_m}^{\text{sp}}]$, $\mathbf{U}_{mm} = \text{diag}[u_1^{\text{sp}}, \ldots, u_{N_m}^{\text{sp}}]$ where $u_m^{\text{sp}} = \sum_{n \in \mathcal{N}_m} \omega_{mn}^{\text{sp}}$, and $\mathbf{W} = [\omega_{mn}^{\text{sp}}]_{m,n=1,\ldots,N_m}$.

This linear system with a Laplacian matrix can be easily solved with conventional linear solvers [62]. Fig. 9 shows examples of our superpixel graph-based propagation.

5.3 Efficient Dense Descriptor on Superpixels

The sampling patterns are transformed with corresponding geometric fields $\mathbf{G}^{\rho}$ and $\mathbf{G}^{\theta}$ as shown in Fig. 10. Specifically, for the mth superpixel $\mathcal{S}_m$, the sampling pattern $(s_{m,l}, t_{m,l}) \in \Lambda_m^{\text{gi-dasc}}$ is transformed from $(s_l, t_l) \in \Lambda^{\text{dasc}}$ with a scale factor $\mathbf{G}_m^{\rho}$ and a rotation factor $\mathbf{G}_m^{\theta}$

$$s_{m,l} = \mathbf{S}_m \mathbf{R}_m s_l, \qquad (18)$$

where the scale matrix $\mathbf{S}_m = \text{diag}[\mathbf{G}_m^{\rho}]$ and the rotation matrix $\mathbf{R}_m$ is defined with rotation $\mathbf{G}_m^{\theta}$. In a similar way, $t_{m,l}$ is also estimated from t_l. Finally, $\Lambda_m^{\text{gi-dasc}}$ is estimated. Furthermore, the patch size N is enlarged as $N\mathbf{G}_m^{\rho}$.

The mth superpixel extended subimage $\mathcal{K}_m$ in Fig. 10a is filtered by a Gaussian filtering with the sigma $\{(\mathbf{G}_m^{\rho})^2 - 0.25\}^{-1/2}$ similar to scale-space theory used in the SIFT [28].

Then, our GI-DASC descriptor $\mathcal{D}_i^{\mathrm{gi-dasc}} = \bigcup_l d_{i,l}^{\mathrm{gi-dasc}}$ for $l = 1, \ldots, L^{\mathrm{gi-dasc}} (= L^{\mathrm{dasc}})$ is encoded with a set of patch similarity between two patches from a transformed sampling pattern $\Lambda_m^{\mathrm{gi-dasc}}$ on each superpixel $\mathcal{S}_m$ such that

$$d_{i,l}^{\mathrm{gi-dasc}} = \mathcal{C}(s_{i,l}, t_{i,l}), \quad (s_{i,l}, t_{i,l}) \in \Lambda_m^{\mathrm{gi-dasc}}, \qquad (19)$$

for $i \in \mathcal{S}_m$. Finally, the dense GI-DASC descriptor is efficiently computed for all the superpixels $\mathcal{S}_m \in \mathcal{S}$. Algorithm 3 summarizes how to compute the GI-DASC descriptor.

Algorithm 3. Geometric-Invariant DASC (GI-DASC)

Input: image f_i, feature detection sampling patterns Λ_i^{det},
  L^{dasc} sampling patterns $(s_{i,l}, t_{i,l}) \in \Lambda_i^{\mathrm{dasc}}$.
Output: the GI-DASC descriptor volume $\mathcal{D}_i^{\mathrm{gi-dasc}}$.
1: Extract feature points $i \in \mathcal{I}'$ with scale ρ_i and rotation θ_i
   using Algorithm 2.
2: Decompose the image f_i into superpixels $\mathcal{S}$.
3: Compute a surface fitting for geometric field $\mathbf{G}_m^{*,\rho}$ and $\mathbf{G}_m^{*,\theta}$ on
   superpixels $\mathcal{S}_m$.
4: Compute a Laplacian matrix $\mathbf{P} + \mu\mathbf{U} - \mu W$ with confidences
   p_m^{sp} and weights $\omega_{mn}^{\mathrm{sp}}$.
5: Compute dense geometric fields $\mathbf{G}_m^{\rho}$ and $\mathbf{G}_m^{\theta}$.
   for $m = 1 : N_m$ **do**
6:   Transform the sampling pattern $\Lambda_m^{\mathrm{dasc}}$ into $\Lambda_m^{\mathrm{gi-dasc}}$.
7:   Compute the GI-DASC descriptor $d_{i,l}^{\mathrm{gi-dasc}} = \mathcal{C}(s_{i,l}, t_{i,l})$
     for $i \in \mathcal{S}_m$ and $(s_{m,l}, t_{m,l}) \in \Lambda_m^{\mathrm{gi-dasc}}$ using Algorithm 1.
   end for

---

6 EXPERIMENTAL RESULTS AND DISCUSSIONS

6.1 Experimental Environments

In experiments, the DASC descriptor was implemented with the following same parameter settings for all datasets: $\{\sigma_c, \tau_c, N, M, L^{\mathrm{dasc}}\} = \{0.5, 0.03, 5 \times 5, 31 \times 31, 128\}$ where M is the support window size, and $\{N_\rho, N_\theta\} = \{4, 36\}$ for candidate sampling patterns. We set the smoothness parameter $\epsilon = 0.03^2$ in the GF [21]. For the GI-DASC, the following parameters were used for all datasets: $\{N_\rho^{\mathrm{wmsd}}, N_\theta^{\mathrm{wmsd}}, N_k, o, \lambda_c, \lambda_p\} = \{3, 12, 4, 10, 0.1, 30\}$. The number of superpixels is set to about 500. We implemented the DASC and GI-DASC descriptor in C++ on Intel Core i7-3770 CPU at 3.40 GHz.

The DASC descriptor was evaluated with other state-of-the-art descriptors, e.g., SIFT [28], DAISY [13], BRIEF [29], LSS [18], and DaLI [25], and other area-based approaches, e.g., ANCC [42], MI+SIFT[1] [40], and RSNCC [9]. We also compared the DASC using a randomized pooling (DASC+RP) with the DASC using a learned randomized pooling (DASC+LRP). Furthermore, the state-of-the-art geometry robust methods such as SID [48], SegSID [48], SegSF [49], GPM [46], DSP [15], and SSF [45] were also compared to the GI-DASC descriptor. For learning the DASC, we built training sets $\mathcal{P}$ from benchmark databases used in each experiment, and these training sets were excluded from experiments.

---

1. For a fair evaluation, we compared only the similarity measure in [40] without further techniques.

(a) Support window size (b) Descriptor dimension
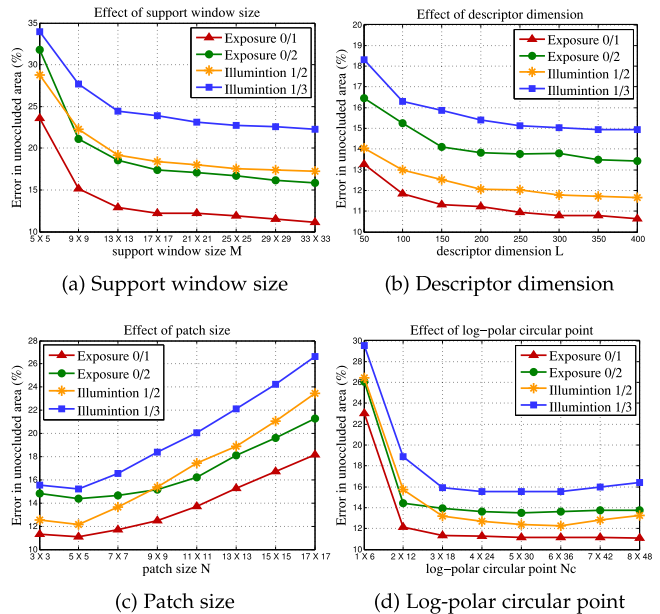
(c) Patch size (d) Log-polar circular point

Fig. 11. Average bad-pixel error rate on Middlebury benchmark [24] of DASC+LRP descriptor with WTA optimization as varying support window size M, descriptor dimension L, patch size N, and log-polar circular point N_c ($\approx N_\rho \times N_\theta$). In each experiment, all other parameters are fixed as initial values in Section 6.1.

6.2 Parameter and Component Analysis

6.2.1 Parameter Sensitivity Analysis

Fig. 11 intensively analyzed the performance of the DASC descriptor as varying associated parameters, including support window size M, descriptor dimension L^{dasc}, patch size N, and the number of log-point circular point N_c. To evaluate the quantitative performance, we measured an average bad-pixel error rate on Middlebury benchmark [24]. The larger the support window size M, the matching quality is improved but the accuracy gain is saturated around 31×31. Using a larger descriptor dimension L^{dasc} yields a better performance since the descriptor encodes more information. Considering the trade-off between efficiency and robustness, $L^{\mathrm{dasc}} = 128$ is set in experiments. When the patch size N increases, the matching quality is degraded since a series of similarity values measured with large patches may lose locally discriminative details. The number of log-polar circular point N_c does not affect the performance much, since optimal patterns can be sampled even from small N_c.

6.2.2 Component-Wise Performance Gain Analysis

The DASC is originally motivated by the LSS concept from [18]. The DASC consists of three key ingredients: adaptive self-correlation (ASC), randomized pooling (RP), and learning sampling pattern. In this context, we analyzed an accuracy gain of the DASC over the LSS on the Middlebury benchmark as shown in Fig. 12. Note that all experiments were done using LSS without max pooling, 'LSS(wo/max)'. The original LSS method [18] uses the SSD for measuring the patch similarity. We replaced the patch similarity of the LSS method with the ASC, named 'LSS(ASC)', and then measured its matching accuracy. As expected, the ASC improves the performance compared to the SSD used in the original LSS. We also evaluated
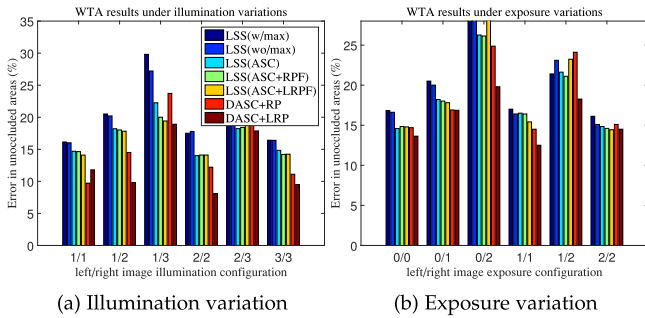
(a) Illumination variation   (b) Exposure variation

Fig. 12. Average bad-pixel error rate for original LSS [18], LSS without max-pooling, LSS with ASC, LSS using randomized-pooling with fixed center pixel, and the DASC descriptor on Middlebury benchmark [24].

the LSS using a randomized pooling with fixed center pixel, 'LSS(ASC+RPF)', and the LSS using a learned randomized pooling with fixed center pixel, 'LSS(ASC+LRPF)'. Unlike center-biased poolings, the DASC chooses sampling patterns randomly ('DASC+RP'), improving the performance. Using learned sampling patterns ('DASC+LRP') also leads to a performance gain.

6.2.3 Edge-Aware Filtering Analysis

In Fig. 13, we analyzed the performance of the DASC descriptor when different EAF is employed for computing $\omega_{i,i'}$. When using a simple, unweighted 'Box' filtering ($\omega_{i,i'} = 1$), the patch similarity (7) becomes a normalized cross-correlation (NCC). In the Box and Gaussian filtering case, there exists a performance limitation. In contrast, all EAF methods show a satisfactory performance, including the bilateral filter [63], the fast bilateral filter [59], the domain transform [56], the fast GF [64], and GF [71]. In experiments, we utilized the GF [71].

6.2.4 WMSD Feature Detector Analysis

In Figs. 14 and 15, we analyzed the feature detection performance of the WMSD detector with a repeatability [66] and recognition rate measure [29] in Mikolajczyk dataset [72]. Compared to conventional feature detection approaches [28], [60], [65], [66], the WMSD detector extracts reliable and distinctive points with a high repeatability thanks to its robustness for modality variations including blur artifacts and illumination changes. Furthermore, compared to conventional gradient-based [28], [67] or intensity-based rotation estimations [68], [69], our WMSD-based rotation

(a) Graffiti   (b) Trees

(c) Bikes   (d) Leuven

Fig. 14. Evaluation of the WMSD detection compared to conventional feature detections, such as SIFT [28], MSER [65], FAST [66], and MSD [60]. The WMSD provides reliable feature detection performance, thus providing reliable hypothesis for initial sparse geometric fields.

estimation combined with the DASC descriptor shows the best performance with a high recognition rate.

6.2.5 Symmetric and Asymmetric Measure Analysis

As shown in Fig. 16, a performance gap between using the asymmetric measure $\tilde{\Psi}(i, j)$ in (9) and the symmetric measure $\Psi(i, j)$ in (7) is negligible, while using the asymmetric measure is much faster.

6.2.6 Runtime Analysis

In Table 1, we compared the computational speed of DASC descriptor with state-of-the-art local descriptors, SIFT [28], DAISY [13], and LSS [18]. The DASC provides state-of-the-art computational speed. It should be noted that through recent more efficient edge-aware filters [64], the runtime of DASC can be further reduced.

6.3 Middlebury Stereo Benchmark

We evaluated our DASC+LRP descriptor compared to other approaches in Middlebury stereo benchmark containing illumination and exposure variations [24]. In experiments, the illumination (or exposure) combination '1/3' indicates

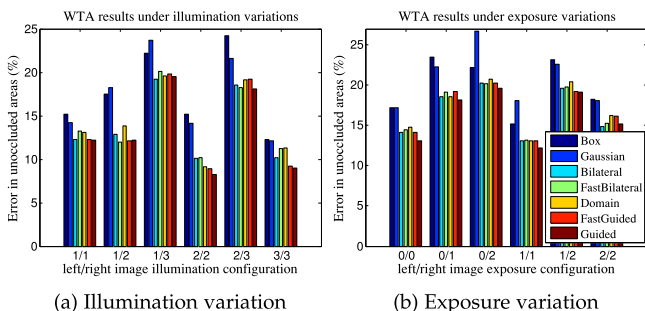(a) Illumination variation   (b) Exposure variation

Fig. 13. Average bad-pixel error rate for the DASC descriptor as varying EAF including Box, Gaussian, Bilateral [63], FastBilateral [55], Domain Transform [56], FastGF [64], and GF [21] on Middlebury benchmark [24].
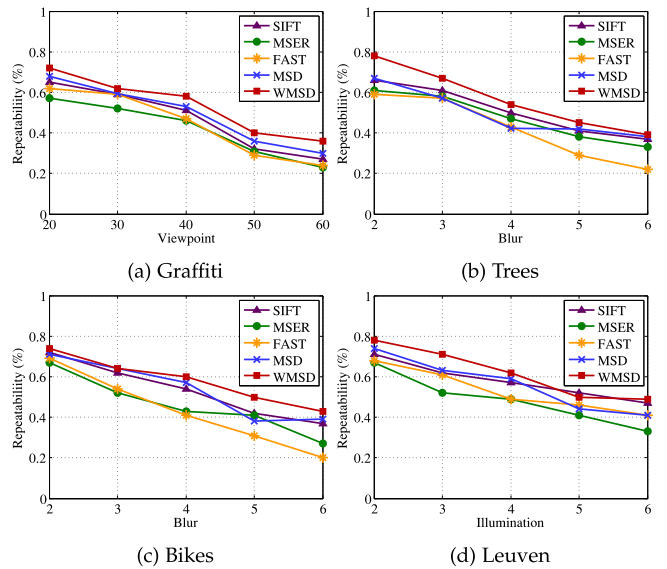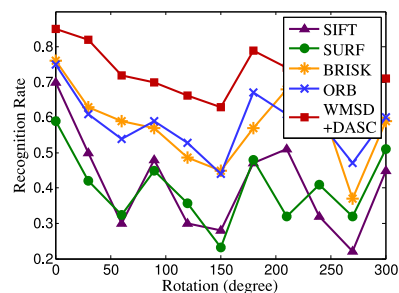

Fig. 15. Evaluation of the WMSD detection compared to conventional rotation estimations. Compared to conventional gradient-based rotation estimation (SIFT [28] and SURF [67]) or intensity-based rotation estimation (BRISK [68] and ORB [69]), our WMSD-based rotation estimation (with the DASC descriptor) shows the best performance.
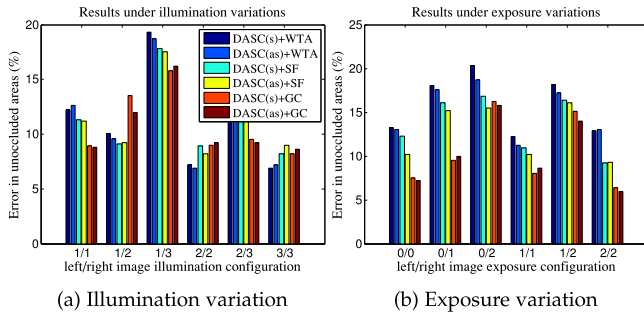
(a) Illumination variation   (b) Exposure variation

Fig. 16. Evaluation of a symmetric measure $\Psi(i,j)$ and an asymmetric measure $\tilde{\Psi}(i,j)$ in the DASC as varying optimization schemes with WTA, SF [14], and GC [70]. It shows that there are no significant performance gaps when using symmetric and asymmetric measure.

TABLE 1
Evaluation of Computational Time

image size	SIFT	DAISY	LSS	DASC†	DASC‡
463×370	130.3 s	2.5 s	31 s	128 s	**1.3 s**
800×600	252 s	3.8	59 s	256 s	**2.1 s**

The brute-force and efficient computation of the DASC is denoted as † and ‡, respectively.

that two images were captured under 1st and 3rd illumination (exposure) conditions, respectively [24]. Fig. 17 shows average bad matching errors in un-occluded areas of depth maps obtained under illumination or exposure variations with the graph-cut (GC) [70] and winner-takes-all (WTA) optimization. Fig. 18 shows disparity maps for severe illumination variations obtained by varying cost functions with the WTA optimization. Our DASC+LRP descriptor achieves the best results both quantitatively and qualitatively. Area-based approaches, e.g., MI+SIFT [40], ANCC [42], and RSNCC [9], are very sensitive to severe radiometric variations, especially when local variations frequently occur. Contrarily, descriptor-based approaches perform better than the area-based approaches. Interestingly, the BRIEF [29] is better than other gradient-based descriptors (SIFT [28] and DAISY [13]) thanks to an ordering robustness.

6.4 Multi-Modal and Multi-Spectral Benchmark

Next, we evaluated our DASC+LRP descriptor with images under modality variations, e.g., RGB-NIR [1], [9], different exposure [8], [9], flash-noflash [8], blurred artifacts [6], [7], and RGB-depth [9]. As varying descriptors and similarity measures, we use the WTA and SIFT flow optimization using the hierarchical dual-layer belief propagation (BP) [14],

(a) Illumination variation   (b) Exposure variation

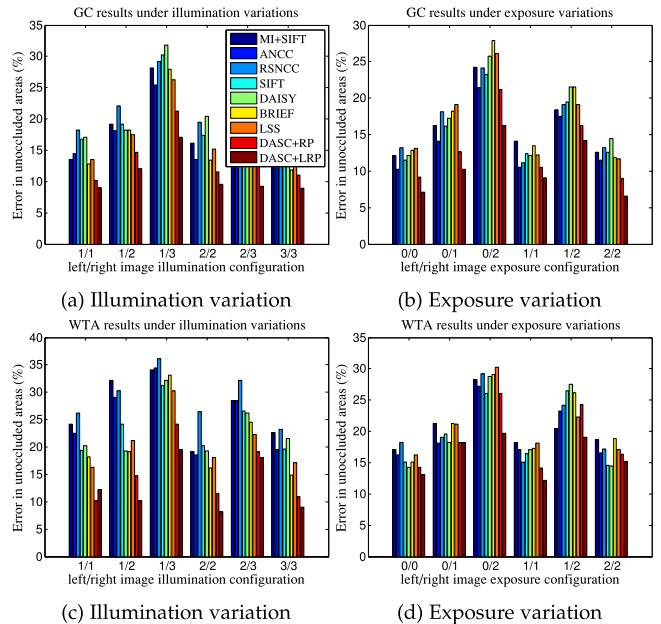(c) Illumination variation   (d) Exposure variation

Fig. 17. Average bad-pixel error rate on Middlebury benchmark with illumination variations and exposure variations. The GC (first row) and WTA (second row) were used for optimization, respectively. Our DASC +LRP shows the best performance with the lowest error rate.

whose code is publicly available. Unlike the Middlebury stereo benchmark, these datasets have no ground truth correspondence maps, and thus we manually obtained ground truth displacements for 100 corner points for all images, and used them for an objective evaluation similar to [9].

Area-based approaches, e.g., MI+SIFT [40], ANCC [42], and RSNCC [9], are very sensitive to local variations. As already described in literatures [9], gradient-based approaches, e.g., SIFT [28] and DAISY [13], have shown limited performance in RGB-NIR pairs where the gradient reversal and inversion frequently appear. The BRIEF [29] cannot deal with noisy and modality varying regions since it considers a pixel difference only. It should be noted that some efforts have been made to estimate reliable flow maps in the motion blur, e.g., blur-flow [73], but they typically employ an iterative matching framework, which relies heavily on an initial estimate. Additionally, they do not scale well to general purpose matching scenarios. Unlike these approaches, the LSS [18] and our descriptor consider the local self-similarities, but the LSS still lacks a discriminative power for dense matching. Our DASC+RP descriptor leveraging patch-wise pooling with adaptive self-correlation provides satisfactory results under modality variations.

(a) Left image   (b) Right image   (c) ANCC [42]   (d) BRIEF [29]   (e) SIFT [28]   (f) LSS [18]   (g) DASC+LRP   (h) Ground Truth
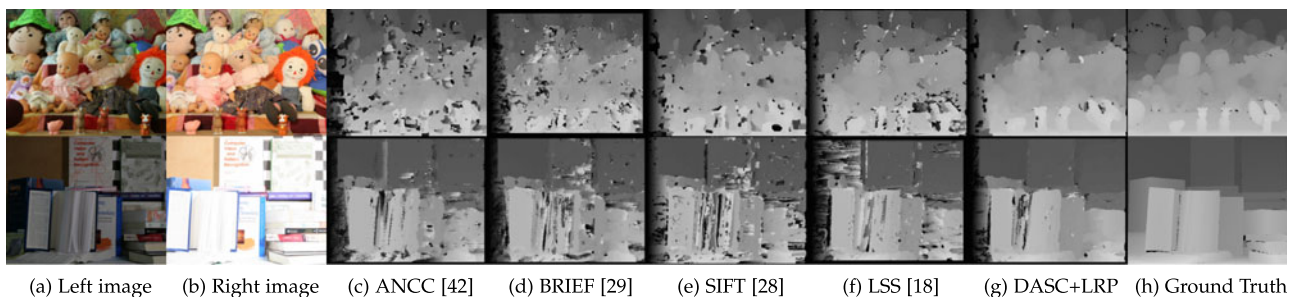
Fig. 18. Comparison of disparity estimation for *Dolls* and *Books* image pairs under illumination combination '1/3' and exposure combination '0/2', respectively. Compared to other approaches, our DASC descriptor estimates accurate and edge-preserved disparity maps while reducing artifacts.

(a) Image pairs   (b) MI+SIFT [40]   (c) BRIEF [29]   (d) DAISY [13]   (e) SIFT [28]   (f) LSS [18]   (g) DASC+RP   (h) DASC+LRP
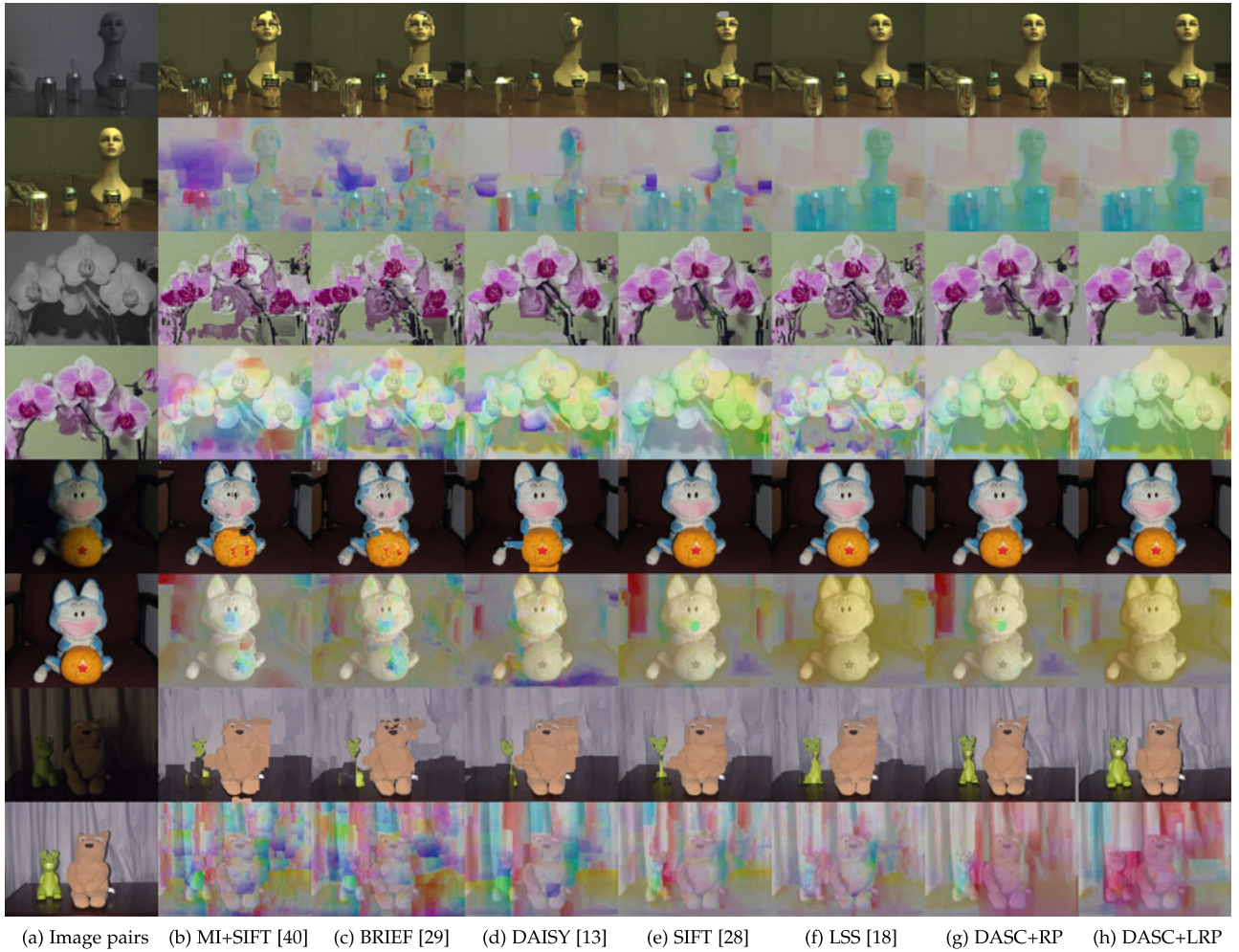
Fig. 19. Comparison of dense correspondence for (from top to bottom) RGB-NIR images and flash-noflash images. The results consist of warped color images and correspondence flow fields overlaid with reference images. Compared to other conventional approaches, our DASC+LRP descriptor estimates reliable dense correspondence fields for challenging multi-modal and multi-spectral image pairs.

By employing the optimal sampling pattern via discriminative learning (DASC+LRP), the matching accuracy was further improved. Figs. 19, 20, and 21 show qualitative evaluation, clearly demonstrating the outstanding performance of our descriptor. Table 2 shows an objective evaluation of DASC+LRP descriptor and other state-of-the-art methods.

6.5 DIML Multi-Modal Benchmark

Since there have been no database with both photometric and geometric variations, we built the DIML multi-modal benchmark [27]. All databases were taken by SONY Cyber-Shot DSC-RX100 camera in a darkroom with the lighting booth GretagMacbeth SpectraLight III. In terms of geometric deformations, we captured 10 geometry image sets by combining geometric variations of viewpoint, scale, and rotation as shown in Fig. 22, and each image set consists of images taken under five different photometric variation pairs including illumination, exposure, flash-noflash, blur, and noise as shown in Fig. 23. Therefore, the DIML multi-modal benchmark consists of 100 images with the size of $1,200 \times 800$. Furthermore, by following [14], we manually built ground truth object annotation maps to evaluate the performance quantitatively, and computed the label transfer accuracy (LTA) $\mathcal{A}^{\text{LTA}}$ such that

$$\mathcal{A}^{\text{LTA}} = \frac{1}{\mathcal{T}_a} \sum_{i \in \mathcal{I}} 1(e_i \neq a_i, a_i > 0), \qquad (20)$$

where the ground-truth annotation is a_i, estimated annotation is e_i, and $\mathcal{T}_a = \sum_{i \in \mathcal{I}} 1(a_i > 0)$ is the number of labeled pixels. This metric has been widely used in wide-baseline matching tasks [15]. Though $\mathcal{A}^{\text{LTA}}$ does not measure a matching performance in a pixel precision, it was shown in [14] that this metric is an excellent alternative enough to evaluate the performance of descriptors in case that there are no ground truth correspondence maps available.

For an image from the reference geometry image set (the first image in Fig. 22), we estimated visual correspondence maps with images from other geometry image set, and then computed the LTA. Furthermore, visual correspondence maps were estimated for each photometric pair. Here, matching results at occluded pixels should be excluded in the evaluation as they have no corresponding pixels. We hence warped an image taken from near into an image taken at a distance, when computing the LTA. The experimental setup for DIML multi-modal benchmark was given in detail at our project page [27].

We compared our two descriptors, DASC and GI-DASC, with conventional descriptors such as SIFT [28], DAISY [13], BRIEF [29], and LSS [18], and state-of-the-arts

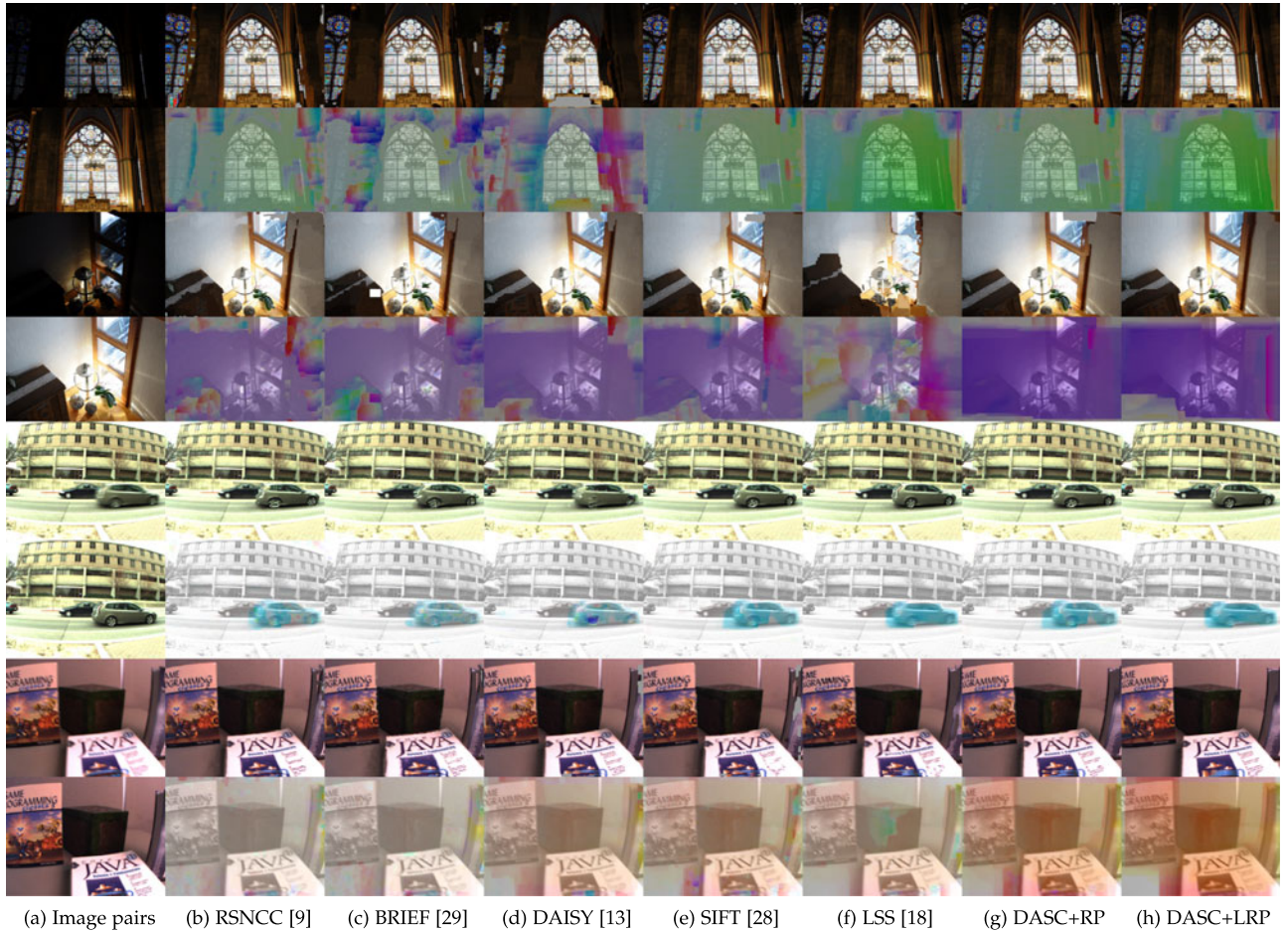| (a) Image pairs | (b) RSNCC [9] | (c) BRIEF [29] | (d) DAISY [13] | (e) SIFT [28] | (f) LSS [18] | (g) DASC+RP | (h) DASC+LRP |

Fig. 20. Comparison of dense correspondence for (from top to bottom) different exposure images and blurred-sharpen images. The results consist of warped color images and correspondence flow fields overlaid with reference images. Compared to other conventional approaches, our DASC+LRP descriptor estimates reliable dense correspondence fields for challenging multi-modal and multi-spectral image pairs.

geometry-invariant approaches such as SID [48], SegSIFT [49], SegSID [49], GPM [46], DSP [15], and SSF [45]. For the sake of simplicity, we omit 'LRP' in the DASC-LRP and GI-DASC-LRP. Fig. 24 shows the LTA error rates as varying photometric and geometric deformations. Fig. 25 shows the average error rates on DIML multi-modal benchmark. Fig. 26 shows qualitative evaluation results. As expected, feature descriptors such as SIFT [28], DAISY [13], BRIEF [29], and LSS [18], though using a powerful global optimization, i.e., hierarchical dual-layer BP [14], exhibit limitations on severe geometric variations, while they provide robustness to some extent for photometric variations. Our DASC descriptor in Fig. 24k shows a better

performance than other descriptors, but it also shows the limitation for severe geometric variations. The GPM [46] had very low performance in terms of flow estimation although it provides plausible warping results. The SID [48] have been proposed to provide geometric robustness, but it is unable to address photometric variations. Segmentation-aware description [49] could improve the matching accuracy of SIFT and SID for geometric variations, but it also has limitation since it also reduces a discriminative power of descriptor itself as shown in Figs. 24g and 24h.

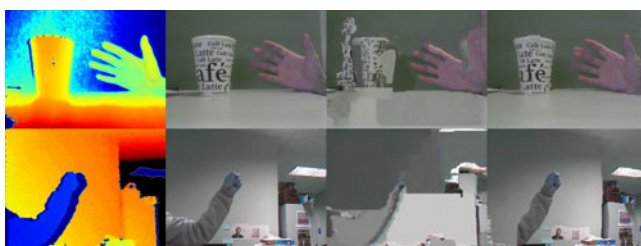| (a) depth | (b) color image | (c) DAISY [13] | (d) DASC+LRP |

Fig. 21. Comparison of dense correspondence for RGB-depth images. The results consist of warped color images.
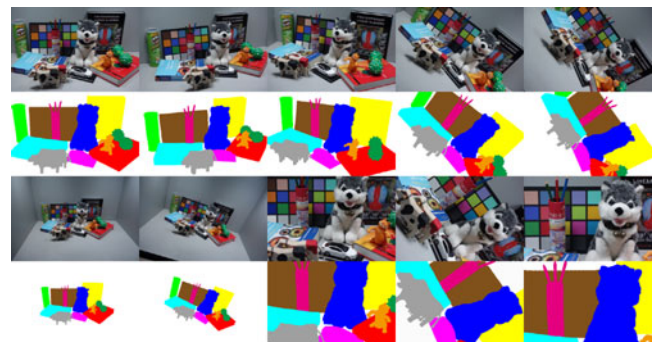

Fig. 22. Examples of DIML multi-modal benchmark. It consists of images taken under 10 different geometric conditions such as viewpoint, scale, rotation, and scale-rotation with ground truth annotation.

TABLE 2
Comparison of Quantitative Evaluation on Multi-Spectral and Multi-Modal Images

	WTA optimization					SF optimization [14]				
	RGB-NIR	flash-noflash	diff. expo.	blur-sharp	Average	RGB-NIR	flash-noflash	diff. expo.	blur-sharp	Average
MI+SIFT [40]	25.13	27.12	28.23	24.21	27.12	17.21	13.24	14.16	20.14	16.87
ANCC [42]	23.21	20.42	25.19	26.14	23.74	18.45	14.14	11.96	19.24	15.94
RSNCC [9]	27.51	25.12	18.21	27.91	24.68	13.41	15.87	9.15	18.21	14.16
SIFT [28]	24.11	18.72	19.42	27.18	22.36	18.51	11.06	14.87	20.78	16.35
DAISY [13]	27.61	26.30	20.72	27.41	25.51	20.42	10.84	12.71	22.91	16.72
BRIEF [29]	29.14	18.29	17.13	26.43	22.75	17.54	9.21	9.54	19.72	14.05
LSS [18]	27.82	19.18	18.21	26.14	22.84	16.14	11.88	9.11	18.51	13.91
DASC+RP	18.21	14.28	12.12	17.11	12.18	15.43	7.51	7.32	12.21	9.68
DASC+LRP	**13.42**	**11.28**	**9.23**	**13.28**	**11.80**	**8.10**	**5.41**	**6.24**	**10.81**	**7.64**

The DSP [15] provides limited performances, since it just uses the SIFT with a fixed scale and rotation. The SSF [45] estimates visual correspondence by repeatedly applying the SIFT on the scale-space while enduring a huge computational complexity, but it still has limitations in terms of computational complexity. Contrarily, the GI-DASC descriptor optimized by hierarchical dual-layer BP [14] provides the robustness for both photometric and geometric deformations as shown in Fig. 24l.

6.6 Non-Rigid Deformation Benchmark

We also evaluated our DASC descriptor compared to other approaches on a recent, publicly available dataset featuring challenging non-rigid deformations and very severe illumination changes [25]. Similar to experiments in Section 6.4, as varying descriptors, we use the the hierarchical dual-layer belief propagation (BP) [14]. Fig. 27 presents dense correspondence estimates for this benchmark [25]. As expected, conventional gradient-based and intensity comparison-based feature descriptors, including SIFT [28], DAISY [13], and BRIEF [29], do not provide reliable correspondence performance. LSS [18] exhibits relatively high performance for illumination changes, but it is vulnerable to non-rigid deformations. Although DaLI [25] provides robust correspondences, it requires considerable computation for dense matching, which is approximately 100 times slower than DASC. Compared to these methods, our DASC descriptor provides a robustness even for non-rigid image deformation, and it is computationally efficient.

6.7 MPI Optical Flow Benchmark

Optical flow methods typically assume only a small displacement between consecutive frames. Several approaches have been proposed to estimate a large displacement

flow vector [74]. However, motion blur and illumination variation can degenerate the performance of these approaches. In order to handle such challenging issues simultaneously, we applied the DASC to the large displacement optical flow (LDOF) approach [74]. It was evaluated on the MPI SINTEL database [11] containing large non-rigid motion as well as specular reflections, motion blur, and defocus blur. The dataset consists of two kind of rendering frames, named clean and final pass, and each set contains 12 sequences with over 500 frames in total [11]. Table 3 shows average end-point error (EPE) results on MPI

(a) DAISY [13]   (b) LSS [18]

(c) SegSIFT [49]   (d) SegSID [49]

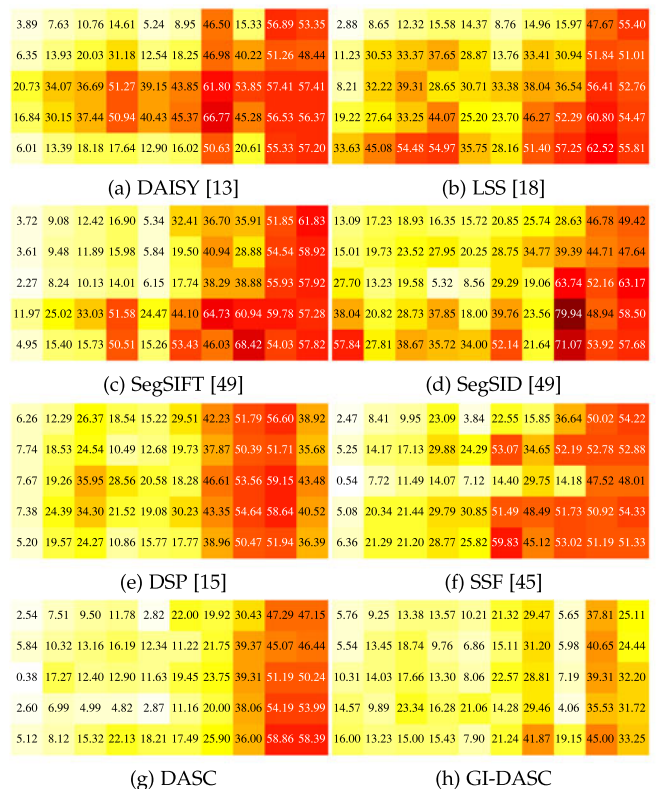(e) DSP [15]   (f) SSF [45]

(g) DASC   (h) GI-DASC

Fig. 24. Comparison of quantitative evaluation on DIML benchmark [27]. Each result represents the LTA for geometric (x-axis) and photometric (y-axis) variations, respectively. The DASC outperforms conventional descriptors such as DAISY [13] and LSS [18]. Interestingly, its accuracy is also higher than those of state-of-the-art geometry-invariant methods including SegSIFT [49], SegSID [49], DSP [15], and SSF [45]. The GI-DASC shows the best performance under varying photometric and geometric conditions. (Best viewed in color.)
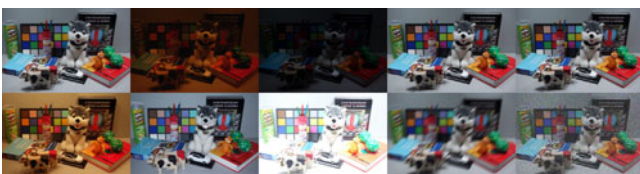

Fig. 23. Examples of DIML multi-modal benchmark. Each geometry image sets in Fig. 22 consists of five different photometric variations such as illumination, exposure, flash-noflash, blur, and noise.
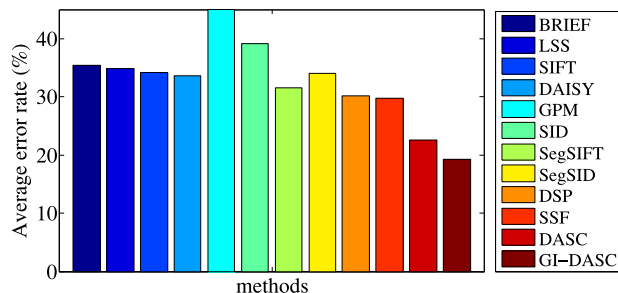
Fig. 25. Average error rates on DIML multi-modal benchmark.

TABLE 3
Comparison of Average EPE on the MPI SINTEL [11]

	Clean Pass		Final Pass	
	all	*unmatched*	*all*	*unmatched*
Classic-NL [12]	7.940	39.821	9.439	43.123
LDOF [74]	7.180	38.124	8.422	42.892
LDOF+BRIEF [29]	6.281	37.841	7.741	41.875
LDOF+LSS [18]	6.182	37.514	7.152	40.332
LDOF+DASC	**5.578**	**36.975**	**6.384**	**38.932**

SINTEL. The DASC achieves a higher gain, compared to other descriptors.

6.8 Limitations

Similar to [22], [44], [45], our GI-DASC approximately determines a relative scale using successive Gaussian smoothing, which might work in only a limited range of scale variation as in Fig. 28. By leveraging an octave structure based on sub-sampling [28], a wider range of scale may be covered.

7 CONCLUSION

The robust novel dense descriptor called the DASC has been proposed for dense multi-spectral and multi-modal correspondences. It leverages an adaptive self-correlation measure and a randomized receptive field pooling learned by linear discriminative learning. Moreover, by making use of fast edge-aware filters, our DASC descriptor is capable of computing the dense descriptor very efficiently. In order to address geometric variations, the GI-DASC descriptor also

(a) Image pairs   (b) DAISY [13]   (c) Seg-SIFT [49]   (d) Seg-SID [48]   (e) DSP [15]   (f) SSF [45]   (g) DASC   (h) GI-DASC
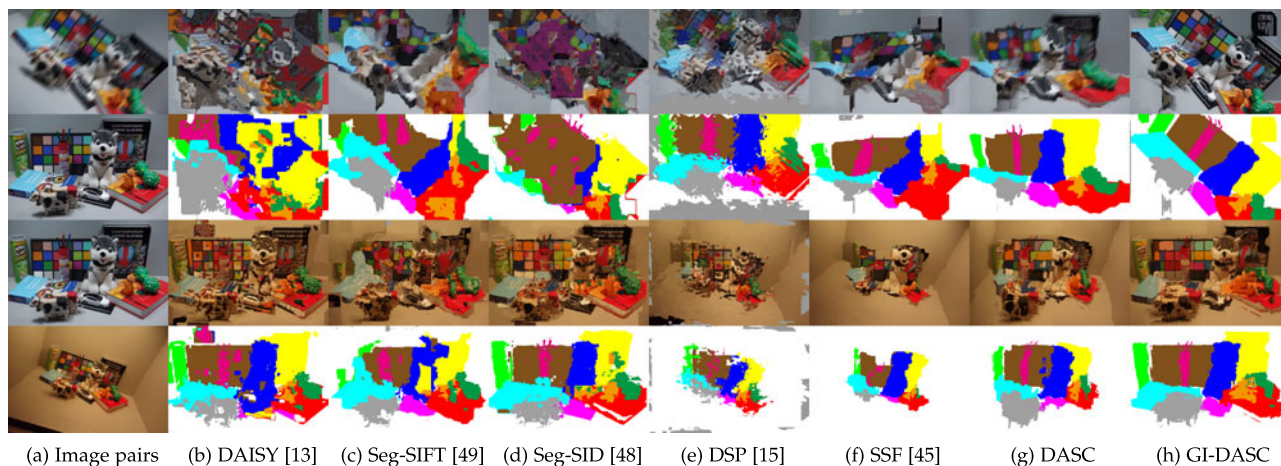
Fig. 26. Comparison of qualitative evaluation on DIML multi-modal benchmark. The results consist of warped color images and warped ground truth annotations. Compared to other conventional descriptors and geometry-invariant approaches, our DASC descriptor estimates reliable dense correspondence fields for image pairs across varying geometric and photometric conditions.

(a) Image pairs   (b) SIFT [28]   (c) DAISY [13]   (d) BRIEF [29]   (e) LSS [18]   (f) DSP [15]   (g) DaLI [25]   (h) DASC
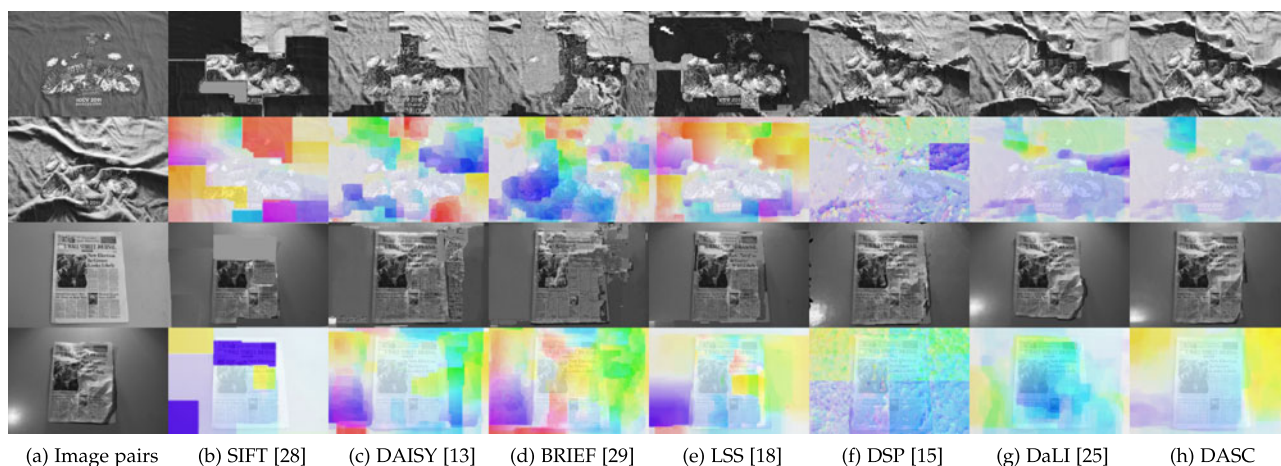
Fig. 27. Comparison of qualitative evaluation on non-rigid deformation benchmark. The results consist of warped color images and correspondence flow fields overlaid with reference images. Compared to other conventional approaches, our DASC descriptor estimates reliable dense correspondence fields for challenging non-rigid deformation image pairs.

| (a) image 1 | (b) image 2 | (c) SSF [45] | (d) GI-DASC |

Fig. 28. Limitations for images under severe geometric variations.

has been proposed by leveraging the efficiency and effectiveness of the DASC through a superpixel-based representation. The DASC and GI-DASC descriptor demonstrated its robustness in establishing dense correspondence between challenging image pairs taken under different modality conditions, e.g., RGB-NIR, different illumination and exposure, flash-noflash, blurring artifacts. We believe our method will serve as an essential tool for several applications using multi-modal and multi-spectral images.

ACKNOWLEDGMENTS

REFERENCES

[1] M. Brown and S. Susstrunk, "Multispectral SIFT for scene category recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 177–184.

[2] Q. Yan, X. Shen, L. Xu, and S. Zhuo, "Cross-field joint image restoration via scale map," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1537–1544.

[3] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

[4] D. Krishnan and R. Fergus, "Dark flash photography," in *Proc. ACM SIGGRAGH*, 2009, Art. no. 96.

[5] G. Petschnigg, M. Agrawals, and H. Hoppe, "Digital photography with flash and no-flash image pairs," in *Proc. ACM SIGGRAGH*, 2004, pp. 664–672.

[6] Y. HaCohen, E. Shechtman, and D. Lishchinski, "Deblurring by example using dense correspondence," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2384–2391.

[7] H. Lee and K. Lee, "Dense 3D reconstruction from severely blurred images using a single moving camera," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 273–280.

[8] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, 2012, Art. no. 203.

[9] X. Shen, L. Xu, Q. Zhang, and J. Jia, "Multi-modal and multi-spectral registration for natural images," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 309–324.

[10] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 7–42, 2002.

[11] D. Butler, J. Wulff, G. Stanley, and M. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.

[12] D. Sun, S. Roth, and M. Black, "Secret of optical flow estimation and their principles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2432–2439.

[13] E. Tola, V. Lepetit, and P. Fua, "DAISY: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, May 2010.

[14] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2368–2382, Dec. 2011.

[15] J. Kim, C. Liu, F. Sha, and K. Grauman, "Deformable spatial pyramid matching for fast dense correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2307–2314.

[16] Y. Li, D. Min, M. S. Brown, M. N. Do, and J. Lu, "SPM-BP: Sped-up patchmatch belief propagation for continuous MRFs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4006–4014.

[17] P. Pinggera, T. Breckon, and H. Bischof, "On cross-spectral stereo matching using dense gradient features," in *Proc. British Mach. Vis. Conf.*, 2012, pp. 1–12.

[18] E. Schechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[19] P. Heinrich, et al., "MIND: Modality indepdent neighbourhood descriptor for multi-modal deformable registration," *Med. Image Anal.*, vol. 16, no. 3, pp. 1423–1435, 2012.

[20] A. Torabi and G. Bilodeau, "Local self-similarity-based registration of human ROIs in pairs of stereo thermal-visible videos," *Pattern Recognit.*, vol. 46, no. 2, pp. 578–589, 2013.

[21] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[22] H. Yang, W. Lin, and J. Lu, "DAISY filter flow: A generalized discrete approach to dense correspondences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3406–3413.

[23] J. Hur, H. Lim, C. Park, and S. C. Ahn, "Generalized deformable spatial pyramid: Geometry-preserving dense correspondence estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1392–1400.

[24] (2002). [Online]. Available: http://vision.middlebury.edu/stereo/

[25] E. Simo-Serra, C. Torras, and F. Moreno-Noguer, "DaLI: Deformation and light invariant descriptor," *Int. J. Comput. Vis.*, vol. 115, no. 2, pp. 136–154, 2015.

[26] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2103–2112.

[27] (2015). [Online]. Available: http://diml.yonsei.ac.kr/~srkim/DASC/

[28] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[29] M. Calonder, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2011.

[30] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 510–517.

[31] S. Saleem and R. Sablatnig, "A robust SIFT descriptor for multi-spectral images," *IEEE Signal Process. Lett.*, vol. 21, no. 4, pp. 400–403, Apr. 2014.

[32] Y. Ye and J. Shan, "A local descriptor based registration method for multispectral remote sensing images with non-linear intensity differences," *ISPRS J. Photogrammetry Remote Sens.*, vol. 90, no. 7, pp. 83–95, 2014.

[33] K. Alex, S. Ilya, and E. H. Geoffrey, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[34] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1573–1585, Aug. 2014.

[35] P. Fischer, A. Dosovitskiy, and T. Brox, "Descriptor matching with convolutional neural networks: A comparison to SIFT," *arXiv:1405.5769*, 2014.

[36] J. Donahue, et al. "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learning*, 2014, pp. 647–655.

[37] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 118–126.

[38] J. Dong and S. Soatto, "Domain-size pooling in local descriptors: DSP-SIFT," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5097–5106.

[39] J. Pluim, J. Maintz, and M. Viergever, "Mutual information based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[40] Y. Heo, K. Lee, and S. Lee, "Joint depth map and color consistency estimation for stereo images with different illuminations and cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1094–1106, May 2013.

[41] J. Xu, Q. Yang, J. Tang, and Z. Feng, "Linear time illumination invariant stereo matching," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 119, pp. 179–193, 2016.

[42] Y. Heo, K. Lee, and S. Lee, "Robust stereo matching using adaptive normalized cross-correlation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 4, pp. 807–822, Apr. 2011.

[43] M. Irani and P. Anandan, "Robust multi-sensor image alignment," in *Proc. 6th Int. Conf. Comput. Vis.*, 1998, pp. 959–966.

[44] T. Hassner, V. Mayzels, and L. Zelnik-Manor, "On SIFTs and their scales," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1522–1528.

[45] W. Qiu, X. Wang, X. Bai, A. Yuille, and Z. Tu, "Scale-space SIFT flow," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 1112–1119.

[46] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, "The generalized patchmatch correspondence algorithm," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 29–43.

[47] 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 *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1854–1861.

[48] I. Kokkinos and A. Yuille, "Scale invariance without scale selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[49] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. M. Noguer, "Dense segmentation-aware descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2890–2897.

[50] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd Eur. Conf. Comput. Vis.*, 1994, pp. 151–158.

[51] B. Fan, Q. Kong, T. Trzcinski, Z. Wang, C. Pan, and P. Fua, "Receptive fields selection for binary feature description," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2583–2595, Jun. 2014.

[52] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[53] C. Lee, A. Bhardwaj, V. Jagadeesh, and R. Piramuthu, "Region-based discriminative feature pooling for scene text recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4050–4057.

[54] T. Trzcinski, M. Christoudias, and V. Lepetit, "Learning image descriptor with boosting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 597–610, Mar. 2015.

[55] Q. Yang, K. Tan, and N. Ahuja, "Real-time O(1) bilateral filtering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 557–564.

[56] E. Gastal and M. Oliveira, "Domain transform for edge-aware image and video processing," in *Proc. ACM SIGGRAGH*, 2011, Art. no. 69.

[57] M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger, "Robust anisotropic diffusion," *IEEE Trans. Image Process.*, vol. 7, no. 3, pp. 421–432, Mar. 1998.

[58] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3017–3024.

[59] S. Paris and F. Durand, "A fast approximation of the bilateral filter using a signal processing approach," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 24–52, 2009.

[60] F. Rombari and L. D. Stefano, "Interest points via maximal self-dissimilarities," in *Proc. 12th Asian Conf. Comput. Vis.*, 2014, pp. 586–600.

[61] S. Kim, B. Ham, S. Ryu, S. J. Kim, and K. Sohn, "Robust stereo matching using probabilistic Laplacian surface propagation," in *Proc. 12th Asian Conf. Comput. Vis.*, 2014, pp. 368–383.

[62] M. Lang, O. Wang, T. Aydic, A. Smolic, and M. Gross, "Practical temporal consistency for image-based graphics applications," *ACM Trans. Graph.*, vol. 31, 2012, Art. no. 34.

[63] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE 6th Int. Conf. Comput. Vis.*, 1998, pp. 839–846.

[64] K. He and J. Sun, "Fast guided filter," *arXiv:1505.00996*, 2015.

[65] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *Proc. British Mach. Vis. Conf.*, 2002, pp. 1–10.

[66] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, Jan. 2010.

[67] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 404–417.

[68] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK : Binary robust invariant scalable keypoints," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2548–2555.

[69] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 2564–2571.

[70] Y. Boykov, O. Yeksler, and R. Zabih, "Fast approximation enermgy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[71] K. He, J. Sun, and X. Tang, "Guided image filtering," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 1–14.

[72] K. Mikolajczyk, et al., "A comparison of affine region detectors," *Int. J. Comput. Vis.*, vol. 65, no. 1, pp. 43–72, 2005.

[73] T. Portz, L. Zhang, and H. Jiang, "Optical flow in the presence of spatially-varying motion blur," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1752–1759.

[74] T. Brox and J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 500–513, Mar. 2011.

Seungryong Kim (S'12) received the BS degree in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2012. He is currently working toward the MS and PhD degrees in electrical and electronic engineering at Yonsei University. His current research interests include computer vision, computational photography, and machine learning, in particular, local and dense feature descriptor, multi-modal image registration, and optimization. He is a student member of the IEEE.

Dongbo Min (M'09-SM'15) received the BS, MS, and PhD degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea, in 2003, 2005, and 2009, respectively. He was with Mitsubishi Electric Research Laboratories, Cambridge, Massachusetts, as a post-doctoral researcher from 2009 to 2010. From 2010 to 2015, he was with Advanced Digital Sciences Center, Singapore. Since 2015, he has been an assistant professor in the Department of Computer Science and Engineering, Chungnam National University, Daejeon, Korea. His current research interests include computer vision, 2D/3D video processing, computational photography, augmented reality, and continuous/discrete optimization. He is a senior member of the IEEE.

Bumsub Ham (M'13) received the BS and PhD degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, Korea, in 2008 and 2013, respectively. From 2014 to 2016, he was a post-doctoral research fellow with Willow Team, INRIA Paris, École Normale Supérieure de Paris, and Centre National de la Recherche Scientifique. He is currently an assistant professor in the School of Electrical and Electronic Engineering, Yonsei University. He received the Grand Prize in Qualcomm Innovation Fellowship in 2012. His current research interests include computer vision, computational photography, and machine learning, in particular, regularization and registration, both in theory and applications. He is a member of the IEEE.

Minh N. Do (M'01-SM'07-F'14) received the BEng degree in computer engineering from the University of Canberra, Canberra, ACT, Australia, in 1997, and the DSc degree in communication systems from Swiss Federal Institute of Technology Lausanne, Lausanne, Switzerland, in 2001. He has been a faculty member with the University of Illinois at Urbana-Champaign, Champaign, Illinois, since 2002, where he is currently a professor in the Department of Electrical and Computer Engineering, and holds joint appointments with the Coordinated Science Laboratory, Beckman Institute for Advanced Science and Technology, and Department of Bioengineering. His research interests include image and multidimensional signal processing, wavelets and multiscale geometric analysis, computational imaging, augmented reality, and visual information representation. He received the Silver Medal from the 32nd International Mathematical Olympiad in 1991, the University Medal from the University of Canberra in 1997, the Doctorate Award from EPFL in 2001, the CAREER Award from the National Science Foundation in 2003, and the Young Author Best Paper Award from the IEEE in 2008. He was named a Beckman Fellow with the Center for Advanced Study, UIUC, in 2006, and received the Xerox Award for Faculty Research from the College of Engineering, UIUC, in 2007. He was a member of the IEEE Signal Processing Theory and Methods Technical Committee and the Image, Video, and Multidimensional Signal Processing Technical Committee, and an associate editor of the *IEEE Transactions on Image Processing*. He is a fellow of the IEEE.

Kwanghoon Sohn (M'92-SM'12) received the BE degree in electronic engineering from Yonsei University, Seoul, Korea, in 1983, the MSEE degree in electrical engineering from the University of Minnesota, Minneapolis, Minnesota, in 1985, and the PhD degree in electrical and computer engineering from North Carolina State University, Raleigh, North Carolina, in 1992. He was a senior member of the research staff with Satellite Communication Division, Electronics and Telecommunications Research Institute, Daejeon, Korea, from 1992 to 1993, and a post-doctoral fellow with MRI Center, Medical School of Georgetown University, Washington, District of Columbia, in 1994. He was a visiting professor with Nanyang Technological University, Singapore, from 2002 to 2003. He is currently a professor in the School of Electrical and Electronic Engineering, Yonsei University. His research interests include 3D image processing, computer vision, and image communication. He is a senior member of the IEEE and a member of the SPIE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.