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 in 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 multi-modal and multi-spectral dense 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 multi-modal and multi-spectral dense 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] and dark flash image [3]. More broadly, flash and no-flash images [4], blurred images [5], [6], and images taken under different radiometric conditions [7] can also be considered as multi-modal [8].

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 [9] or optical flow [10], [11] 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 [13], [14], [15], an optimizer itself cannot address an inherent limitation without suitable matching descriptors [16].

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 \(^{[17]}\). 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 \(^{[18]}, [19]}\), 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 measure, which 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 \(^{[20]}\).

Furthermore, in order to address geometric variation problems such as the scale and rotation, 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 key-points 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 for dense correspondence \(^{[21], [22]}\), 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 \(^{[23]}\), (2) multi-modal and multi-spectral dataset including RGB-NIR images \(^{[1], [8]}\), different exposure \(^{[7]}\), flash-noflash images \(^{[7]}\), and blurry images \(^{[5], [6]}\), and (3) MPI optical flow benchmark containing specular reflections, motion blur, and defocus blur \(^{[10]}\). 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 \(^{[24]}\). 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 \(^{[25]}\).

2 RELATED WORK

2.1 Feature Descriptors

As a pioneering work, the scale invariant feature transform (SIFT) was first introduced by Lowe \(^{[26]}\) 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) \(^{[27]}\) and fast retina keypoint (FREAK) \(^{[28]}\), have been proposed. Unlike these sparse descriptors, Tola et al. developed a dense descriptor, called DAISY \(^{[12]}\), 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 de-
veloped [29], but these gradient-based descriptors have an inherent limitation similar to the SIFT, especially when an image gradient varies across different modality images. Schechtman and Irani introduced the LSS descriptor [17] 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) [19]. The LSS also has been applied to the registration of multi-spectral remote sensing images [30]. For multi-modal medical image registration, Heinrich et al. proposed a modality independent neighborhood descriptor (MIND) [18] 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) [31] for estimating correspondences. For designing explicit, discriminative feature descriptors, intermediate activations from CNN architecture are extracted [32], [33], [34], [35], 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 [35], [36]. 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 [37], 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 [38], 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 [39]. Although cross-correlation based methods such as an adaptive normalized cross-correlation (ANCC) [40] 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 [41], but it also shows a limitation for general image matching tasks. A robust selective normalized cross-correlation (RSNCC) [8] 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) [13] optimization, many methods have been proposed to alleviate geometric variation problems, including deformable spatial pyramid (DSP) [14], scale-less SIFT flow (SLS) [42], scale-space SIFT flow (SSF) [43], and generalized DSP (GDSP) [22]. 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) [44] was proposed for efficient matching leveraging a randomized search scheme. The DAISY Filter Flow (DFF) [21], which exploits DAISY descriptor [12] with PatchMatch Filter (PMF) [45], was proposed to provide geometric invariance. However, their weak spatial smoothness often induces mismatched results. The scale invariant descriptor (SID) [46] was proposed to encode geometric robustness on the descriptor itself, but it is not tailored to multi-modal matching. Segmentation-aware approach [47] was proposed to provide geometric robustness for descriptors, e.g., SIFT [26] or SID [46], but it may have a negative effect on the discriminative power of the descriptor.

### 3 Background

Let us define an image as \( f_i : \mathcal{I} \rightarrow \mathbb{R} \) for pixel \( i \), where \( \mathcal{I} \subset \mathbb{N}^2 \) is a discrete image domain. Given the image \( f_i \), a dense descriptor \( D_i : \mathcal{I} \rightarrow \mathbb{R}^L \) is defined on a local support window \( \mathcal{R}_i \) centered at pixel \( i \) with a feature dimension \( L \). Conventionally, 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 [26]) or an intensity comparison (BRIEF [27]) cannot capture coherent matching evidences, resulting erroneous local minima in estimating dense correspondences.

Unlike these conventional descriptors, the LSS descriptor \( D_i^{\text{lss}} \) measures a correlation between two patches \( F_i \) and \( F_j \) centered at two pixels \( i \) and \( j \) within a local support window \( \mathcal{R}_i \) [17]. As shown in Fig. 3(a), 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, \( D_i^{\text{lss}} = \bigcup_{l=1}^{L^{\text{lss}}} d_i^{\text{lss},l} \) for \( l = 1, \ldots, L^{\text{lss}} \) is a \( L^{\text{lss}} \times 1 \) feature vector, and \( d_i^{\text{lss},l} \) can be computed as follows:

\[
d_{i,l}^{\text{lss}} = \max_{j \in \text{bin}(l)} \{ C(i,j) \},
\]
where \( \text{bin}_i(l) = \{ j \mid j \in \mathcal{R}_i, \rho_{r-1} < |i - j| \leq \rho_r, \theta_{a-1} < \angle(i - j) < \theta_a \} \) with a log radius \( \rho_r \) for \( r \in \{ 1, \cdots, N_p \} \) and a quantized angle \( \theta_a \) for \( a \in \{ 1, \cdots, N_\theta \} \) with \( \rho_0 = 0 \) and \( \theta_0 = 0 \). In that case, \( \mathcal{B} = N_p \times N_\theta \). 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 \( \sigma_s \):

\[
\mathcal{C}(i, j) = \exp \left( -\text{SSD}(\mathcal{F}_i, \mathcal{F}_j) / \sigma_s \right).
\]

This LSS descriptor has been shown to be robust in cross-domain object detection [17], 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 \( \text{bin}_i(l) \) 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. 3(a), 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 [17]. 3) From the relationship between Census transform [48] and BRIEF [27] 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 \( \mathcal{P}_i \) as shown in Fig. 3(b). It is defined as \( \mathcal{P}_i = \{ j \mid j \in \mathcal{R}_i, |i - j| = \rho_r, \angle(i - j) = \theta_a \} \) where the number of points is defined as \( N_c = N_p \times N_\theta + 1 \) and has a higher density of points near a center pixel, similar to DAISY descriptor [12].

Given \( N_c \) points in \( \mathcal{P}_i \), there exist \( N_{pc} = \binom{N_c - 1}{2} \) 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 \( \mathcal{L}^{\text{dasc}} \) sampling patterns from \( N_{pc} \) pattern candidates. Our descriptor \( \mathcal{D}^{\text{dasc}} = \bigcup \mathcal{L}^{\text{dasc}} \) for \( l = 1, \cdots, \mathcal{L}^{\text{dasc}} \) is encoded with a set of patch similarity between two patches based on sampling patterns that are selected from \( \mathcal{P}_i \):

\[
d^{\text{dasc}}_{h,i,l} = \mathcal{C}(s_{l,i}, t_{l,i}), \quad s_{l,i}, t_{l,i} \in \mathcal{P}_i,
\]

where \( s_{l,i} \) and \( t_{l,i} \) are \( l \)-th 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 \( t_{l,i} - s_{l,i} \) for \( l = 1, \cdots, \mathcal{L}^{\text{dasc}} \), enabling a fast computation of dense descriptors, which will be detailed in Sec. 4.3. Although the DASC descriptor uses only sparse

Fig. 4. Visualization of patch-wise receptive fields of the DASC descriptor learned from the training set \( \mathcal{P} \) built with the Middlebury benchmark [23], multi-modal benchmark [8], and the MPI SINTEL benchmark [10]. Similar to [19], we stack all patch-wise receptive fields learned from each training image, and normalized them with the maximal value.

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 [49], we employ a discriminative learning to obtain optimal sampling patterns within a local support window. Given candidate sampling patterns \( \Lambda_i = \{(s_{l,i}, t_{l,i}) | l = 1, \cdots, N_{pc} \} \), our goal is to select the best sampling patterns which describe two support pairs, allowing the descriptor to consider the majority of pixels in the support window and reflect original image attributes effectively.

Our approach exploits support vector machines (SVMs) with a linear kernel [50]. For learning, we build a dataset \( \mathcal{D} = \{ (\mathcal{R}_h, \mathcal{R}_t, y_h) | h = 1, \cdots, N_{tr} \} \) where \( (\mathcal{R}_h, \mathcal{R}_t)^\mathcal{L} \) 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 \( \mathcal{D} \) was built with images captured under varying illumination conditions and/or with imaging devices [8, 10, 23]. In experiments, \( N_{tr} = 10,000 \).

First, the feature \( r_h = \bigcup_i r_{h,i,l} \) that describes two support window pairs \( \mathcal{R}_h^1 \) and \( \mathcal{R}_t^2 \) is defined

\[
r_{h,i,l} = \exp \left( -\left( d_{h,l}^{\text{dasc},1} - d_{h,l}^{\text{dasc},2} \right)^2 / 2\sigma_r^2 \right),
\]

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

\[
\mathcal{Q}(r_h) = v^Tr_h + b,
\]

where the weight \( v = \bigcup_i v_i \) indicates an amount of contribution of each candidate sampling pattern, and \( b \) is a bias. Learning \( v \) can be formulated as minimizing

\[
\mathcal{E}_{\text{svm}}(v) = \|v\|^2 + C_{\text{svm}} \sum_{h=1}^{N_{tr}} l_{\text{hinge}}(y_h \cdot \mathcal{Q}(r_h)),
\]

where the hinge loss function \( l_{\text{hinge}}(x) = \max(0, 1 - x) \) and \( C_{\text{svm}} \) represents a regularization parameter. We use LIBSVM
4.2 Adaptive Self-Correlation Measure

With estimated sampling patterns \((s_i,t_i)\), the DASC descriptor measures a patch similarity using an adaptive self-correlation (ASC) measure in order to robustly encode a local internal layout of selfsimilarities. 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 \((s,t) \in \Lambda_{\text{dasc}}\), the adaptive self-correlation \(\Psi(s,t)\) between two patches \(F_s\) and \(F_t\) centered at pixels \(s\) and \(t\) is computed as follows:

\[
\Psi(s,t) = \frac{\sum_{s',t'} \omega_{s,s'} \omega_{t,t'} (f_{s'} - G_s)(f_{t'} - G_t)}{\sqrt{\sum_{s'} \left\{ \omega_{s,s'} (f_{s'} - G_s)^2 \right\}} \sqrt{\sum_{t'} \left\{ \omega_{t,t'} (f_{t'} - G_t)^2 \right\}}},
\]

(7)

where \(s',t' \in F_s\) and \(t' \in F_t\) and weighted averages on \(F_s\) and \(F_t\) are defined as \(G_s = \sum_{s'} \omega_{s,s'} f_{s'}\) and \(G_t = \sum_{t'} \omega_{t,t'} f_{t'}\).

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 [20], [52], [53]. 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 [40], 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 \(F_s\) and \(F_t\) is computed with a truncated exponential function, which has been widely used in the robust estimator [55]:

\[
C(s,t) = \max(\exp(-(1 - |\Psi(s,t)|)/\sigma_c), \tau_c),
\]

(8)

where \(\sigma_c\) is a bandwidth of Gaussian kernel and \(\tau_c\) is a truncation parameter. Here, a absolute value of \(\Psi(s,t)\) is used to mitigate the effect of intensity reverses. The correlation \(C(s_i,t_i)\) for \(i\) is normalized with an unit norm for all \(l\).

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 \(C(s_i,t_i)\) for all patch pairs belonging to \((s_i,t_i) \in \Lambda_{\text{dasc}}\) 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 \((s_i,t_i)\) efficiently, we employ a constant-time edge-aware filter (EAF), e.g., the guided filter (GF) [20]. 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 \(F_s\) 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 [56]. 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 Sec. 6.2.3. For efficient descriptor, we also re-arrange the sampling pattern \((s_i,t_i)\) to referenced-biased pairs \((i,j) = (i,i + t_i - s_i)\). (7) is then approximated as follows:

\[
\tilde{\Psi}(i,j) = \frac{\sum_{i',j'} \omega_{i,i'} (f_{i'} - G_{i'}) (f_{j'} - G_{j'})}{\sqrt{\sum_{i'} \omega_{i,i'} (f_{i'} - G_{i'})^2} \sqrt{\sum_{j'} \omega_{i,i'} (f_{j'} - G_{j'})^2}}.
\]

(9)
Algorithm 1: Dense Adaptive Self-Correlation (DASC)

**Input**: image \( f_i \), candidate sampling patterns \( \Lambda_l \), training patch pairs dataset \( P \). 

**Output**: the DASC descriptor volume \( D_{\text{dasc}} \).

/\* Offline Procedure */
1: Compute \( r_l \) using (4) for possible candidate sampling patterns \( \Lambda_l \) on training support window pairs \( P \).
2: Learn a weight \( \omega \) by optimizing (5).
3: Select the maximal \( \tilde{D}_{\text{dasc}} \) sampling patterns \((s_l, t_l)\) in terms of weight \( \omega \). Let \( \tilde{\Lambda}_{\text{dasc}} \) be the set of sampling patterns obtained.

/\* Online Procedure */
4: Compute \( \tilde{D}_l = \sum_{i'} \omega_{i',i} f_{i'} \) for all pixel \( i \).
5: Compute \( \tilde{G}_l = \sum_{i'} \omega_{i',i} f_{i'}^2 \).
6: for \( l = 1 : \tilde{\Lambda}_{\text{dasc}} \) do
   6.1: Re-arrange \((s_l, t_l)\) in \( \tilde{\Lambda}_{\text{dasc}} \) as \((i, j) = (i, i + t_l - s_l)\).
   6.2: Compute \( \tilde{G}_{ij} = \sum_{i',j'} \omega_{i',i} \tilde{f}_{i',j'} \).
   6.3: Compute \( \tilde{G}_{ij}^2 = \sum_{i',j'} \omega_{i',i} \tilde{f}_{i',j'}^2 \).
   6.4: Compute \( \tilde{G}_{ij,j} = \sum_{i',j'} \omega_{i',i} \tilde{f}_{i',j'} \).
   6.5: Compute \( \tilde{G}_{ij,j} = \sum_{i',j'} \omega_{i',i} \tilde{f}_{i',j'} \).
10: Compute the DASC descriptor \( D_{\text{dasc}} \) by re-indexing sampling patterns such that \( d_{l,ij} = \tilde{C}(s_l, t_l, i, j) \).
end for

where \( G_l = \sum_{i'} \omega_{i',i} f_{i'} \). Furthermore, \( G_{ij} = \sum_{i',j'} \omega_{i',i} \tilde{f}_{i',j'} \) which means weighted average of \( f_{i'} \in F_j \) with a guidance image \( f_{i'} \in F_i \). It is worth noting that the robustness of \( \Psi(s,t) \) can be still applied to \( \Psi(i,j) \) since their difference is just weight factors.

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

\[
\frac{G_{ij,j} - G_{i,j} \cdot G_{i,j}}{\sqrt{G_{ij} - G_{i,j}^2 \cdot G_{ij,j}^2}},
\]

where \( G_{ij} = \sum_{i'} v_{i',j} f_{i'} \), \( G_{i,j} = \sum_{i',j'} v_{i',i} \tilde{f}_{i',j'} \), and \( G_{ij,j} = \sum_{i',j'} v_{i',i} \tilde{f}_{i',j'} \).

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 \( \omega_{s,t'} \) of \( \Psi(i,j) \) in (9). The symmetric weight case in the DASC can also be computed similar to Sec. 4.3. After re-arranging the sampling pattern as \((i,j) = (i, i + t_l - s_l)\), the (9) can be then decomposed as similar in (10)

\[
\Psi(i,j) = \frac{G_{i,j,j} - G_{i,j,j} - G_{i,j} \cdot G_{i,j}}{\sqrt{G_{i,j} - G_{i,j}^2 \cdot G_{i,j,j}^2}},
\]

where \( G_{i,j,j} = \sum_{i',j'} \omega_{i',i} \omega_{j,j'} f_{i',j'}^2 \), \( G_{i,j,j} = \sum_{i',j'} \omega_{i',i} \tilde{f}_{i',j'} \), and \( G_{i,j,j} = \sum_{i',j'} \omega_{i',i} \tilde{f}_{i',j'} \).

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 \( O(IL) \) for \( L < L \).

5 GEOMETRY-INvariant DASC descriptor

Similar to the DAISY (12), 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 [20] can be used for 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 [26], 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 sensitive to multi-modal and multi-spectral deformation. In Conventional feature detectors, e.g., SIFT [26], are very sensitive to multi-modal and multi-spectral deformation. In order to infer dense geometric fields from sparse geometric fields \( \rho_i \) and initial sparse geometric fields \( G^*_i \), we construct a histogram with angles \( \{ \theta_k \} \) such that

\[
\Omega^*_i = \{ \theta_k \} \quad \text{and} \quad \rho_i = \sum_{\theta_k} \omega_i \theta_k
\]

is defined with \( \rho_i \) and initial sparse geometric fields \( G^*_i \). Similar to the DASC, (12) can be computed efficiently using constant time EAF [20, 57].

We extract the index set \( \Pi^o_i \) for the \( o \) most smallest value \( \Psi^w_{i,l}^{\text{mod},k} \) for all \( l \), i.e., \( o \) nearest neighbors for center patch in Fig. 8(b). It should be noted that parameter \( o \) trades distinctiveness and computational efficiency [58]. We then compute feature response map \( \Omega^*_i \) by estimating the summation of \( \Phi^k(i,l) \) for \( l \in \Pi^o_i \) such that

\[
\Omega^*_i = \sum_{l \in \Pi^o_i} \Phi^k(i,l).
\]

For feature response maps \( \Omega_i = \{ \Omega^*_i \} \), the local maxima are obtained by the non maximal suppression, which compares \( \Omega^*_i \) to its 8 neighbors on the current scale and 18 neighbors on the \((k+1)^{th}\) and \((k-1)^{th}\) scales. Similar to SIFT [26], a feature point \( i \in I' \) is detected only if \( \Omega^*_i \) has an extreme value compared to all of these neighbors, and its scale \( \rho_i \) is defined with \( \rho_k \), where \( I' \subset I \) is a sparse discrete image domain.

A canonical orientation is further associated to \( i \in I' \) by constructing a histogram with angles \( \angle(t_{i,1} - s_{i,1}) \) for \( l \in \Pi^o_i \) weighted by \( \Phi^k(i,l) \) as

\[
l_{\text{hist}}(i, \theta) = \sum_{l \in \Pi^o_i} \Phi^k(i,l) \cdot \delta(\angle(t_{i,1} - s_{i,1}) - \theta),
\]

where \( \delta \) is the Kronecker delta function. Then, we simply choose the direction corresponding to the highest bin in the histogram, i.e., \( \theta_k = \text{argmax}_{\theta_k} l_{\text{hist}}(i, \theta) \). The WMSD detector is summarized in Algorithm 2.

5.2 Superpixel Graph-Based Propagation

In order to infer dense geometric fields from sparse geometric fields \( \rho_i \) and \( G^*_i \) for \( i \in I' \), we decompose the image \( f \) as superpixel \( S = \{ S_m \} \), where \( S_m \cap S_n = \emptyset \) and \( m \neq n \), \( S_m \cap S_n = \emptyset \), \( m = 1, \ldots, N_m \), where \( N_m \) is the number of superpixels. The geometric fields \( G^p_m \) and \( G^\theta_m \) are fitted on each superpixel \( S_m \) as the average of sparse geometric fields \( \rho_i \) and \( G^*_i \) for \( i \in I' \) is defined. Note that this fitting operation is performed only when \( I' \cap S_m \) exists, i.e., the superpixel includes sparse feature points (at least, 1). Finally, the \( G^{*p} = \bigcup_m G^{*p}_m \in \mathbb{R}^{N_m} \) and \( G^{*\theta} = \bigcup_m G^{*\theta}_m \) are constructed for all superpixels.

Similar to [59], our approach then formulates an inference of dense geometric fields \( G^p \) and \( G^\theta \) as a constrained optimization problem where surface-fitted sparse geometric fields \( G^{*p} \) and \( G^{*\theta} \) are interpreted as soft constraints. For the sake of simplicity, we omit \( \rho \) and \( \theta \) since they can be computed using the same method. The energy function of our superpixel-based propagation is defined as follows:

\[
\sum_m \left\{ p^{mp}_m (G_m - G^*_m)^2 + \mu \sum_{n \in N_m} \omega^{mp}_{mn} (G_m - G_n)^2 \right\},
\]

where \( \mu \) is a regularization parameter. Here, the first term encodes the dissimilarity between final geometric fields \( G_m \) and initial sparse geometric fields \( G^*_m \). \( p^{mp}_m \) is an index...
Compute the GI-DASC descriptor solved with conventional linear solvers \[60\]. Fig. 9 shows geometric fields (scale \(G\), rotation \(G^\rho\)) in (e)-(h) are propagated into dense geometric fields (scale \(G^\rho\), rotation \(G^\theta\)) in (i)-(l).

function, which is 1 for valid (constraint) superpixel, and 0 otherwise. The second term imposes the constraint that two adjacent superpixels \(m\) and \(n\) may have similar geometric fields according to superpixel feature affinity \(\omega_{mn}\), 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 \(\nu_m^c\) 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 \(\nu_m^c \in \mathbb{R}^{18}\). For an NIR image, appearance feature is defined on 1-channel intensity domain such that \(\nu_m^c \in \mathbb{R}^2\). 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 \(\nu_m^p \in \mathbb{R}^2\) is defined as a spatial centroid coordinate within superpixels. Based on these superpixel features, a superpixel feature affinity \(\omega_{mn}^{sp}\) between two adjacent superpixel \(m\) and \(n\) in \(\mathcal{N}_m\) is computed as

\[
\omega_{mn}^{sp} = \exp(-||\nu_m^c - \nu_n^c||^2/\lambda_c - ||\nu_m^p - \nu_n^p||^2/\lambda_p),
\]

where \(\lambda_c\) and \(\lambda_p\) 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

\[
(P + \mu U - \mu W)G = PG^*, \tag{17}
\]

where \(P_{mm} = \text{diag}[p_{1m}^{sp}, ..., p_{Nm}^{sp}], U_{mm} = \text{diag}[u_{1m}^{sp}, ..., u_{Nm}^{sp}]\), \(u_{nm}^{sp} = \sum_{n \in \mathcal{N}_m} \omega_{nm}^{sp}\), and \(W = [\omega_{mn}^{sp}]_{m,n = 1, \ldots, N_m}\).

This linear system with a Laplacian matrix can be easily solved with conventional linear solvers \[60\]. Fig. 9 shows examples of our superpixel graph-based propagation.

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

**Input**: image \(I_t\), feature detection sampling patterns \(\Lambda_{\text{dasc}}^{\text{sp}}\), \(\Lambda_{\text{dasc}}^{\text{sp}}\) sampling patterns \((s_{t1}, t_{t1}) \in \Lambda_{\text{dasc}}^{\text{sp}}\).

**Output**: the GI-DASC descriptor volume \(\Phi_{\text{dasc}}^{\text{sp}}\).

1. Extract feature points \(i \in \mathcal{T}\) with scale \(\rho_i\) and rotation \(\theta_i\) using Algorithm 2.
2. Decompose the image \(I_t\) into superpixels \(\mathcal{S}\).
3. Compute a surface fitting for geometric field \(G_{\text{dasc}}^{sp, \rho}\) and \(G_{\text{dasc}}^{\theta}\) on superpixels \(\mathcal{S}_m\).
4. Compute a Laplacian matrix \(P + \mu U - \mu W\) with confidences \(p_{il}^{sp}\) and weights \(\omega_{il}^{sp}\).
5. Compute dense geometric fields \(G_{\text{dasc}}^{\rho}\) and \(G_{\text{dasc}}^{\theta}\) for \(m = 1 : N_m\)
6. Transform the sampling pattern \(\Lambda_{\text{dasc}}^{\rho}\) into \(\Lambda_{\text{dasc}}^{\rho} \ominus\) for \(i \in \mathcal{S}_m\) and \((s_{t1}, t_{t1}) \in \Lambda_{\text{dasc}}^{\rho} \ominus\) using Algorithm 1.

5.3 Efficient Dense Descriptor on Superpixels

The sampling patterns are transformed with corresponding geometric fields \(G^\rho\) and \(G^\theta\) as shown in Fig. 10. Specifically, for the \(m\)-th superpixel \(S_m\), the sampling pattern \((s_{m1}, t_{m1}) \in \Lambda_{\text{dasc}}^{\rho} \ominus\) is transformed from \((s_{t1}, t_{t1}) \in \Lambda_{\text{dasc}}\) with a scale factor \(G_{\text{dasc}}^{\rho}\) and a rotation factor \(G_{\text{dasc}}^{\theta}\),

\[
s_{m1} = S_m R_{m1} s_{t1}, \tag{18}
\]

where the scale matrix \(S_m = \text{diag}[G_{\text{dasc}}^{\rho}]\) and the rotation matrix \(R_{m1}\) is defined with rotation \(G_{\text{dasc}}^{\theta}\). In a similar way, \(t_{m1}\) is estimated from \(t_{t1}\). Finally, \(\Lambda_{\text{dasc}}^{\rho} \ominus\) is estimated. Furthermore, the patch size \(N\) is enlarged as \(NG_{\text{dasc}}^{\rho}\).

The \(m\)-th superpixel extended subimage \(K_m\) in Fig. 10(a) is filtered by a Gaussian filtering with the sigma \((G_{\text{dasc}}^{\rho})^2 - 0.25)^{1/2}\) similar to scale-space theory used in the SIFT \[26\]. Then, our GI-DASC descriptor \(\Phi_{\text{dasc}}^{\rho} \ominus\) is encoded with a set of patch similarity between two patches from a transformed sampling pattern \(\Lambda_{\text{dasc}}^{\rho} \ominus\) on each superpixel \(S_m\) such that

\[
d_{il}^{\rho} = C(s_{il}, t_{il}), \quad (s_{il}, t_{il}) \in \Lambda_{\text{dasc}}^{\rho} \ominus, \tag{19}
\]

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

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:
For a fair evaluation, we compared only the similarity measure in [17] without further techniques.
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.

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.

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

In Fig. 14 and Fig. 15 we analyzed the feature detection performance of the WMSD detector with a repeatability measure [54] and recognition rate measure [27] in Mikolajczyk et al. [64]. In experiments, we utilized the GF [69].
(a) Image pairs 
(b) RSNCC [8] 
(c) BRIEF [27] 
(d) DAISY [12] 
(e) SIFT [26] 
(f) LSS [17] 
(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.

6.2.5 Symmetric and asymmetric measure analysis

As shown in Fig. 16, a performance gap between using the asymmetric measure $\Psi(i, j)$ 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 [26], DAISY [12], and LSS [17]. The DASC provides state-of-the-art computational speed. It should be noted that through recent more efficient edge-aware filters [62], 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 [23]. In experiments, the illumination (or exposure) combination ‘1/3’ indicates that two images were captured under 1st and 3rd illumination (exposure) conditions, respectively [23]. 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) [68] 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 [38], ANCC [40], and RSNCC [8], 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 [27] is better than other gradient-based descriptors (SIFT [26] and DAISY [12]) 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], 6 different exposure [7], flash-no-flash [7], and blurred artifacts [5]. As varying descriptors and similarity measures, we use the WTA and SIFT flow optimization using the hierarchical dual-layer belief propagation (BP) [13], 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 displacement vectors for 100 corner points for all images, and used them for an objective evaluation similar to [8].

Area-based approaches, e.g., MI+SIFT [38], ANCC [40], and RSNCC [8], are very sensitive to local variations. As already described in literatures [8], gradient-based approaches, e.g., SIFT [26] and DAISY [12], have shown limited performance in RGB-NIR pairs where the gradient reversal and inversion frequently appear. The BRIEF [27] 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 [71], 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 [17] 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 matching. Our DASC+RP descriptor considers the local self-similarities, but the LSS still lacks a discriminative power for dense matching.
TABLE 2

Comparison of quantitative evaluation on multi-spectral and multi-modal images.

|          | WTA optimization       | SF optimization      |
|----------|------------------------|----------------------|
|          | RGB-NIR | flash-noflash | diff. expo. | blur-sharp | Average | RGB-NIR | flash-noflash | diff. expo. | blur-sharp | Average |
| MI+SIFT  | 25.13   | 27.12       | 28.19       | 24.21      | 27.12   | 17.21   | 13.24      | 14.16      | 20.14      | 16.87   |
| ANCC     | 23.21   | 20.42       | 26.14       | 19.45      | 23.74   | 18.45   | 14.14      | 11.96      | 19.24      | 15.94   |
| RSNCC    | 27.51   | 25.12       | 18.21       | 27.91      | 24.68   | 13.41   | 15.87      | 9.15       | 18.21      | 14.16   |
| SIFT     | 24.11   | 18.72       | 19.42       | 27.18      | 22.36   | 18.51   | 11.06      | 20.78      | 16.35      |         |
| DAISY    | 27.61   | 26.30       | 20.72       | 27.41      | 25.51   | 20.42   | 10.84      | 12.71      | 22.91      | 16.72   |
| BRIEF    | 29.14   | 18.29       | 17.13       | 26.43      | 22.75   | 17.54   | 9.21       | 9.54       | 19.72      | 14.05   |
| LSS [17] | 27.82   | 19.18       | 18.21       | 26.14      | 22.84   | 16.14   | 11.88      | 9.91       | 18.51      | 13.91   |
| DASC+RP  | 13.42   | 11.28       | 9.23        | 13.28      | 11.80   | 8.10    | 5.41       | 6.24       | 10.81      | 7.64    |
| DASC+LRP |         |             |             |            |         |         |           |            |            |         |

Fig. 21. 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.

Fig. 22. Examples of DIML multi-modal benchmark. Each geometry image set in Fig. 21 consists of 5 different photometric variations such as illumination, exposure, flash-noflash, blur, and noise.

combining geometric variations of viewpoint, scale, and rotation as shown in Fig. 21 and each image set consists of images taken under 5 different photometric variation pairs including illumination, exposure, flash-noflash, blur, and noise as shown in Fig. 22. Therefore, the DIML multi-modal benchmark consists of 100 images with the size of 1200 x 800. Furthermore, by following [13], we manually built ground truth object annotation maps to evaluate the performance quantitatively, and computed the label transfer accuracy (LTA) \( A_{LTA} \) such that

\[
A_{LTA} = \frac{1}{T_a} \sum_{i \in I} 1(e_i \neq a_i, a_i > 0)
\]

where the ground-truth annotation is \( a_i \), estimated annotation is \( e_i \), and \( T_a = \sum_{i \in I} 1(a_i > 0) \) is the number of labeled pixels. This metric has been widely used in wide-baseline matching tasks [14]. Though \( A_{LTA} \) does not measure a matching performance in a pixel wise, it was shown in [13] 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. 21), 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 [25].

We compared our two descriptors, DASC and GI-DASC, with conventional descriptors such as SIFT, DAISY, BRIEF, LSS, and state-of-the-art geometry-invariant approaches such as SIFD, SegSIFT, SegSISD, GPM, DSP, and SSF. For the sake of simplicity, we omit 'LRP' in GI-DASC-LRP and GI-DASC-LRP.
Fig. 23 shows the LTA error rates as varying photometric and geometric deformations. Fig. 24 shows qualitative evaluation results. As expected, feature descriptors such as SIFT [26], DAISY [12], BRIEF [27], and LSS [17], though using a powerful global optimization, i.e., hierarchical dual-layer BP [13], exhibit limitations on severe geometric variations, while they provide robustness to some extent for photometric variations. Our DASC descriptor in Fig. 23(k) shows a better performance than other descriptors, but it also shows the limitation for severe geometric variations. The GPM [44] had very low performance in terms of flow estimation although it provides plausible warping results. The SID [46] have been proposed to provide geometric robustness, but it is unable to address photometric variations. Segmentation-aware description [47] 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 Fig. 23(g) and (h). The DSP [14] provides limited performances, since it just uses the SIFT with a fixed scale and rotation. The SSF [43] 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 [13] provides the robustness for both photometric and geometric deformations as shown in Fig. 23(l). Fig. 25 shows the average error rates on DIML multi-modal benchmark.

### 6.6 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 [72]. 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 [72]. It was evaluated on the MPI Sintel database [10] 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 [10]. Table 3 shows average end-point error (EPE) results on MPI Sintel. The DASC achieves a higher gain, compared to other descriptors.

### 6.7 Limitations

Similar to [21], [42], [43], 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. 26. By leveraging an octave structure based on sub-sampling [26], a wider range of scale may be covered.
Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2013R1A2A2A01068338).

References

[1] M. Brown and S. Susstrunk, “Multispectral sift for scene category recognition,” In Proc. of CVPR, 2011.
[2] Q. Yan, X. Shen, L. Xu, and S. Zhou, “Cross-field joint image restoration via scale map,” In Proc. of ICCV, 2013.
[3] D. Krishnan and R. Fergus, “Dark flash photography,” In Proc. of ACM SIGGRAPH, 2009.
[4] G. Petchenik, M. Agrawals, and H. Hoppe, “Digital photography with flash and no-flash image pairs,” In Proc. of ACM SIGGRAPH, 2004.
[5] Y. HaCohen, E. Shechtman, and E. Lischinsky, “Deblurring by example using dense correspondence,” In Proc. of ICCV, 2013.
[6] H. Lee and K. Lee, “Dense 3d reconstruction from severely blurred images using a single moving camera,” In Proc. of CVPR, 2013.
[7] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, “Robust patch-based hdr reconstruction of dynamic scenes,” In Proc. of ACM SIGGRAPH, 2012.
[8] X. Shen, L. Xu, Q. Zhang, and J. Jia, “Multi-modal and multi-spectral registration for natural images,” In Proc. of ECCV, 2014.
[9] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” IJCV, vol. 47, no. 1, pp. 7–42, 2002.
[10] D. Butler, J. Wulf, G. Stanley, and M. Black, “A naturalistic open source movie for optical flow evaluation,” In Proc. of ECCV, 2012.
[11] D. Sun, S. Roth, and M. Black, “Secret of optical flow estimation and their principles,” In Proc. of CVPR, 2010.
[12] E. Tola, V. Lepetit, and P. Fua, “Daisy: An efficient dense descriptor applied to wide-baseline stereo,” IEEE Trans. PAMI, vol. 32, no. 5, pp. 815–830, 2010.
[13] C. Liu, J. Yuen, and A. Torralba, “Nonparametric scene parsing via label transfer,” IEEE Trans. PAMI, vol. 33, no. 12, pp. 2368–2382, 2011.
[14] J. Kim, C. Liu, F. Sha, and K. Grauman, “Deformable spatial pyramid matching for fast dense correspondences,” In Proc. of CVPR, 2013.
[15] Y. Li, D. Min, M. S. Brown, M. N. Do, and J. Lu, “Spm-bp: Speed-up patchmatch belief propagation for continuous mrfs,” In Proc. of ICCV, 2015.
[16] P. Pinngtera, T. Breckon, and H. Bischof, “On cross-spectral stereo matching using dense gradient features,” In Proc. of BMVC, 2012.
[17] E. Schechtman and M. Irani, “Matching local self-similarities across images and videos,” In Proc. of CVPR, 2007.
[18] P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, V. Gleeson, S. Brady, and A. Schnabel, “Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration,” Med. Image Anal., vol. 16, no. 3, pp. 1423–1435, 2012.
[19] A. Torabi and G. Bilodeau, “Local self-similarity-based registration of human rois in pairs of stereo thermal-visible videos,” Pattern Recognition, vol. 46, no. 2, pp. 578–589, 2013.
[20] K. He, J. Sun, and X. Tang, “Guided image filtering,” IEEE Trans. PAMI, vol. 35, no. 6, pp. 1397–1409, 2013.
[21] J. Yang, W. Lin, and J. Lu, “Daisy filter flow: A generalized discrete approach to dense correspondences,” In Proc. of CVPR, 2014.
[22] J. Hur, H. Lim, C. Park, and S. C. Ahn, “Generalized deformable spatial pyramid: Geometry-preserving dense correspondence estimation,” In Proc. of CVPR, 2015.
[23] http://vision.middlebury.edu/stereo/
[24] D. Sun, S. Roth, and M. Black, “Decaf: A deep convolutional activation feature for generic visual recognition,” In Proc. of IJCV, 2014.
[25] E. Shechtman and M. Irani, “Matching local self-similarities for multi-modal remote sensing images with non-linear intensity differences,” ISPRS J. Photogram. Remote Sens., vol. 90, no. 7, pp. 83–95, 2014.
[26] K. Alex, S. Iyia, and E. H. Geoffrey, “Imagenet classification with deep convolutional neural networks,” In Proc. of NIPS, 2012.
[27] K. Simonyan, A. Vedaldi, and A. Zisserman, “Learning local feature descriptors using convex optimisation,” IEEE Trans. PAMI, vol. 36, no. 8, pp. 1573–1585, 2014.
[28] P. Fischer, A. Dosovitskiy, and T. Brox, “Descriptor matching with convolutional neural networks: A comparison to sift,” arXiv:1405.5769, 2014.
[29] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” In Proc. of IJCV, 2014.
[30] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, “Discriminative learning of deep convolu-
tional feature point descriptors,” In Proc. of IJCV, 2015.
[31] J. Dong and S. Soatto, “Domain-size pooling in local descriptors: Des-p-sift,” In Proc. of CVPR, 2015.
[32] J. Pluim, J. Maintz, and M. Viergever, “Mutual information based registration of medical images: A survey,” IEEE Trans. MI, vol. 22, no. 8, pp. 986–1004, 2003.
[33] Y. Heo, K. Lee, and S. Lee, “Joint depth map and color consistency estimation for stereo images with different illuminations and cameras,” IEEE Trans. PAMI, vol. 35, no. 5, pp. 1094–1106, 2013.
[34] J. Xu, Q. Yang, J. Tang, and Z. Feng, “Linear time illumination invariant stereo matching,” IJCV, 2016.
[35] Y. Heo, K. Lee, and S. Lee, “Robust stereo matching using adaptive normalized cross-correlation,” IEEE Trans. PAMI, vol. 33, no. 4, pp. 807–822, 2011.
[36] M. Irani and P. Anandan, “Robust multi-sensor image alignment,” IEEE Trans. of IC, 1998.
[37] T. Hassner, V. Mayzles, and L. Zelnik-Manor, “On sifts and their scales,” In Proc. of CVPR, 2012.
[38] W. Qiu, X. Wang, X. Bai, A. Yuille, and Z. Tu, “Space-scale sift flow,” In Proc. of WACV, 2014.
[39] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein, “The generalized patchmatch correspondence algorithm,” In Proc. of ECCV, 2010.
[40] 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. of CVPR, 2013.
[41] I. Kokkinos and A. Yuille, “Scale invariance without scale selection,” In Proc. of CVPR, 2008.
[42] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. M. Noguer, “Dense segmentation-aware descriptors,” In Proc. of CVPR, 2013.
[43] R. Zabih and J. Woodfill, “Non-parametric local transforms for computing visual correspondence,” In Proc. of ECCV, 1994.
[44] B. Fan, Q. Kong, T. Trzcinski, and Z. Wang, “Receptive fields selection for binary feature description,” IEEE Trans. IP, vol. 23, no. 6, pp. 2583–2595, 2014.
[50] C. Chang and C. Lin, “Libsvm: A library for support vector machines,” ACM Trans. IST, vol. 2, no. 3, pp. 1–27, 2011.
[51] C. Lee, A. Bhardwaj, V. Jagadeesh, and R. Piramuthu, “Region-based discriminative feature pooling for scene text recognition,” In Proc. of CVPR, 2014.
[52] T. Trzcinski, M. Christoudias, and V. Lepetit, “Learning image descriptor with boosting,” IEEE Trans. PAMI, vol. 37, no. 3, pp. 597–610, 2015.
[53] Q. Yang, K. Tan, and N. Ahuja, “Real-time o(1) bilateral filtering,” In Proc. of CVPR, 2009.
[54] E. Gastal and M. Oliveira, “Domain transform for edge-aware image and video processing,” In Proc. of ACM SIGGRAPH, 2011.
[55] M. J. Black, G. Sapiro, D. H. Marimont, and D. Heeger, “Robust anisotropic diffusion,” IEEE Trans. IP, vol. 7, no. 3, pp. 421–432, 1998.
[56] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, “Fast cost-volume filtering for visual correspondence and beyond,” In Proc. of CVPR, 2011.
[57] S. Paris and F. Durand, “A fast approximation of the bilateral filter using a signal processing approach,” IJCV, vol. 81, no. 1, pp. 24–52, 2009.
[58] F. Rombari and L. D. Stefano, “Interest points via maxiaml self-dissimilarities,” In Proc. of ACCV, 2014.
[59] S. Kim, B. Ham, S. Ryu, S. J. Kim, and K. Sohn, “Robust stereo matching using probabilistic laplacian surface propagation,” In Proc. of ACCV, 2014.
[60] M. Lang, O. Wang, T. Aydic, A. Smolic, and M. Gross, “Practical temporal consistency for image-based graphics applications,” In Proc. of SIGGRAPH, 2012.
[61] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” In Proc. of ICCV, 1998.
[62] K. He and J. Sun, “Fast guided filter,” arXiv:1505.00996, 2015.
[63] J. Matas, O. Chum, and T. Urban, M. amd Pajdla, “Robust wide baseline stereo from maximally stable extremal regions,” In Proc. of BMVC, 2002.
[64] E. Rosten, R. Porter, and T. Drummond, “Faster and better: a machine learning approach to corner detection,” IEEE Trans. PAMI, vol. 32, no. 1, pp. 105–119, 2010.
[65] H. Bay, T. Tuytelaars, and L. V. Gool, “Surf: Speeded up robust features,” In Proc. of ECCV, 2006.
[66] S. Leutenegger, M. Chli, and R. Siegwart, “Brisk : Binary robust invariant scalable keypoints,” In Proc. of ICCV, 2011.
[67] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: an eficient alternative to sift or surf,” In Proc. of ICCV, 2011.
[68] Y. Boykov, O. Yeksler, and R. Zabih, “Fast approximation energy minimization via graph cuts,” IEEE Trans. PAMI, vol. 23, no. 11, pp. 1222–1239, 2001.
[69] K. He, J. Sun, and X. Tang, “Guided image filtering,” In Proc. of ECCV, 2010.
[70] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, “A comparison of affine region detectors,” IJCV, vol. 65, no. 1, pp. 43–72, 2005.
[71] T. Portz, L. Zhang, and H. Jiang, “Optical flow in the presence of spatially-varying motion blur,” In Proc. of CVPR, 2012.
[72] T. Brox and J. Malik, “Large displacement optical flow: Descriptor matching in variational motion estimation,” IEEE Trans. PAMI, vol. 33, no. 3, pp. 500–513, 2011.