DASC: Dense Adaptive Self-Correlation Descriptor for Multi-modal and Multi-spectral Correspondence

Seungryong Kim[1], Dongbo Min[2,3], Bumsub Ham[4], Seungchul Ryu[1], Minh N. Do[5], Kwanghoon Sohn[1]
[1]Yonsei University [2]Chungnam Nat. University [3]ADSC [4]Inria [5]UIUC
http://seungryong.github.io/DASC/

In this supplemental materials, we provide more detailed analysis and results for the DASC descriptor.

- In Section 1, we describe how Eq. (10) is derived from Eq. (9).
- In Section 2, we provide additional experimental results to evaluate the accuracy and runtime efficiency of the DASC descriptor when using the symmetric weight in Eq. (7) and the asymmetric weight in Eq. (9), respectively.
- In Section 3, we show the multi-modal and multi-spectral dataset used in the sampling pattern learning for the patch-wise receptive field pooling, and visualize the estimated sampling pattern.
- In Section 4, we analyze the effect of two parameters (local support window size and feature dimension) used in the DASC descriptor, and provide more results in three datasets; Middlebury stereo benchmark, multi-modal and multi-spectral image pairs, and MPI SINTEL optical flow benchmark.

1. Derivation of Decomposition Eq. (10) from Eq. (9)

In this section, we describe the derivation of Eq. (10) from Eq. (9). By using an asymmetric weight $\omega_{i,i'}$ in adaptive self-correlation $\tilde{\Psi}(i,j)$, we can decompose the adaptive self-correlation into several weighted sum operations. This enables us to further reduce the computational complexity required for computing the DASC descriptor.

$$\begin{aligned}
\tilde{\Psi}(i,j) &= \frac{\sum_{i',j'} \omega_{i,i'}(f_{i'} - \mathcal{G}_i)(f_{j'} - \mathcal{G}_{i,j})}{\sqrt{\sum_{i'} \omega_{i,i'}(f_{i'} - \mathcal{G}_i)^2}\sqrt{\sum_{i',j'} \omega_{i,i'}(f_{j'} - \mathcal{G}_{i,j})^2}} \\
&= \frac{\sum_{i',j'} \omega_{i,i'} f_{i'} f_{j'} - \mathcal{G}_{i,j} \sum_{i'} \omega_{i,i'} f_{i'} - \mathcal{G}_i \sum_{i',j'} \omega_{i,i'} f_{j'} + \mathcal{G}_i \mathcal{G}_{i,j}}{\sqrt{\sum_{i'} \omega_{i,i'} f_{i'}^2 - 2\mathcal{G}_i \sum_{i'} \omega_{i,i'} f_{i'} + \mathcal{G}_i^2}\sqrt{\sum_{i',j'} \omega_{i,i'} f_{j'}^2 - 2\mathcal{G}_{i,j} \sum_{i',j'} \omega_{i,i'} f_{j'} + \mathcal{G}_{i,j}^2}} \\
&= \frac{\mathcal{G}_{i,ij} - \mathcal{G}_{i,j}\mathcal{G}_i - \mathcal{G}_i\mathcal{G}_{i,j} + \mathcal{G}_i\mathcal{G}_{i,j}}{\sqrt{\mathcal{G}_{i^2} - 2\mathcal{G}_i\mathcal{G}_i + \mathcal{G}_i^2}\sqrt{\mathcal{G}_{i,j^2} - 2\mathcal{G}_{i,j}\mathcal{G}_{i,j} + \mathcal{G}_{i,j}^2}} \\
&= \frac{\mathcal{G}_{i,ij} - \mathcal{G}_i \cdot \mathcal{G}_{i,j}}{\sqrt{\mathcal{G}_{i^2} - {\mathcal{G}_i}^2} \cdot \sqrt{\mathcal{G}_{i,j^2} - {\mathcal{G}_{i,j}}^2}},
\end{aligned}$$

where $\mathcal{G}_i = \sum_{i'} \omega_{i,i'} f_{i'}$, $\mathcal{G}_{i,j} = \sum_{i',j'} \omega_{i,i'} f_{j'}$, $\mathcal{G}_{i,ij} = \sum_{i',j'} \omega_{i,i'} f_{i'} f_{j'}$, $\mathcal{G}_{i^2} = \sum_{i'} \omega_{i,i'} f_{i'}^2$ and $\mathcal{G}_{i,j^2} = \sum_{i',j'} \omega_{i,i'} f_{j'}^2$.

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

2. Symmetric Weights $\omega_{s,s'}\omega_{t,t'}$ vs. Asymmetric Weights $\omega_{s,s'}$ in the DASC descriptor

This section analyzes the performance of the DASC descriptor when using a symmetric weight $\omega_{s,s'}\omega_{t,t'}$ in Eq. (7) and with an asymmetric weight $\omega_{s,s'}$ in Eq. (9). For using the symmetric weight, we first re-arrange the sampling pattern $(s_{i,l}, t_{i,l})$ to reference-biased pairs $(i, j) = (i, i + t_{i,l} - s_{i,l})$. With $\bar{f}_{i'} = f_{i'} - \mathcal{G}_i$ and $\bar{f}_{j'} = f_{j'} - \mathcal{G}_{i,j}$, Eq. (7) can be re-written as

$$\Psi(i,j) = \frac{\sum_{i',j'} \omega_{i,i'}\omega_{j,j'}\bar{f}_{i'}\bar{f}_{j'}}{\sqrt{\sum_{i'} \omega_{i,i'}^2 \bar{f}_{i'}^2}\sqrt{\sum_{j'} \omega_{j,j'}^2 \bar{f}_{j'}^2}}.$$

After computing $\bar{f}_{i'}$ and $\bar{f}_{j'}$ for all pixels, the numerator is computed for each sampling pattern offset (i, j) as weighted average of $\bar{f}_{i'}\bar{f}_{j'}$ with symmetric weights $\omega_{i,i'}\omega_{j,j'}$. The denominator is also computed as weighted average of $\bar{f}_{i'}^2$ with weights $\omega_{i,i'}^2$ and weighted average of $\bar{f}_{j'}^2$ with weights $\omega_{j,j'}^2$. In experiments, we used the guided filter (GF) [10], which computes the weighted average in a constant time $O(I)$ (I: image size).

However, the symmetric weight $\omega_{s,s'}\omega_{t,t'}$ varies for each $l \in L$ where L is the number of sampling patterns, and thus this weight should be repeatedly computed for each l, resulting in a huge complexity for computing the numerator in Eq. (7). Moreover, $\omega_{s,s'}\omega_{t,t'}$, $\omega_{s,s'}^2$, and $\omega_{t,t'}^2$ should be computed with a range distance using 6-D vector (or 2D vector), when an input is a color image (or 2-D vector), thus significantly increasing the runtime of constant time edge-aware filtering algorithms (e.g., GF [10]).

To alleviate this problem, our approach employs only the asymmetric weight $\omega_{s,s'}$ for accelerating the computation of the adaptive self-correlation. We found that this modification does not degenerate the performance of the descriptor severely. Fig. 1 and 2 shows the disparity maps obtained using the DASC descriptor with symmetric and asymmetric weights. Fig. 3 shows average bad-pixel error rates on the Middlebury benchmark [1]. A performance gap between using the asymmetric weight and the symmetric weight is negligible.

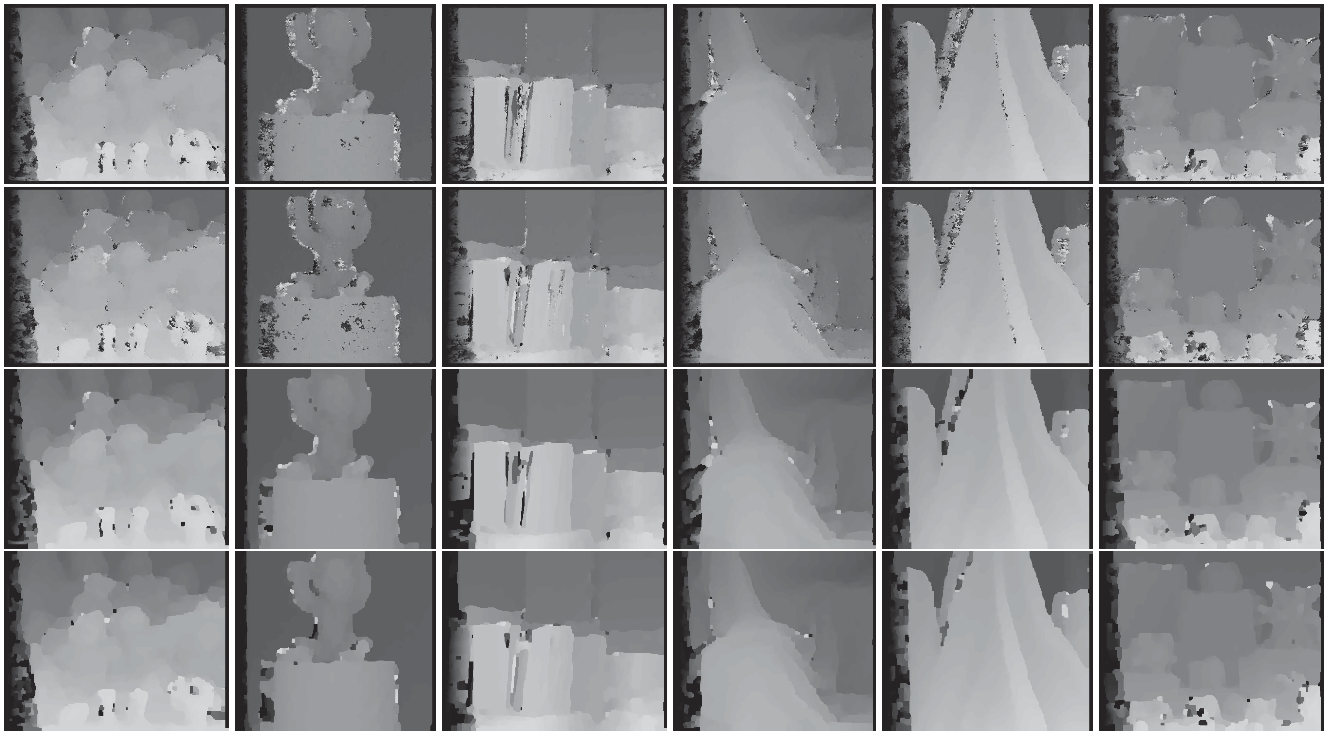
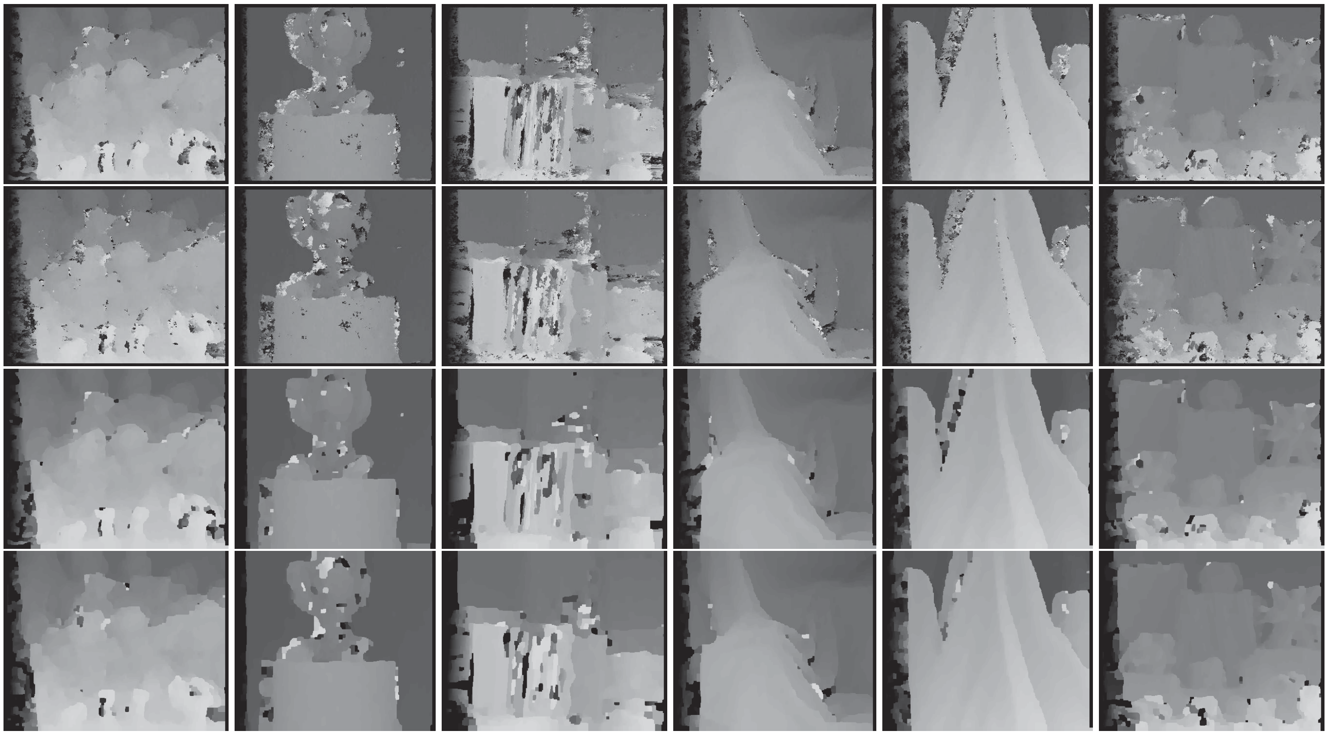

(a) Dolls (b) Baby1 (c) Books (d) Cloth3 (e) Cloth4 (f) Moebius

Figure 1. Comparison of the disparity estimation for *Dolls*, *Baby1*, *Books*, *Cloth3*, *Cloth4*, and *Moebius* image pairs taken under exposure combination ‘0/0’. The first two rows shows the disparity maps obtained using the DASC descriptor with asymmetric and symmetric weights, where the winner-takes-all (WTA) method is used for optimization. The third and fourth rows shows the disparity maps, where the Graph Cuts (GC) method is used for optimization.



(a) Dolls (b) Baby1 (c) Books (d) Cloth3 (e) Cloth4 (f) Moebius

Figure 2. Comparison of the disparity estimation for *Dolls*, *Baby1*, *Books*, *Cloth3*, *Cloth4*, and *Moebius* image pairs taken under exposure combination '0/2'. The first two rows shows the disparity maps obtained using the DASC descriptor with asymmetric and symmetric weights, where the winner-takes-all (WTA) method is used for optimization. The third and fourth rows shows the disparity maps, where the Graph Cuts (GC) method is used for optimization.

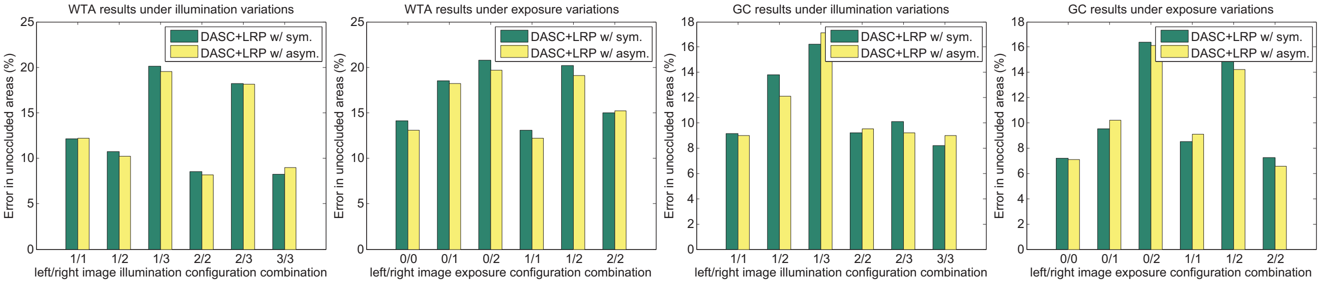


Figure 3. Comparison of average bad-pixel error rates on Middlebury benchmark for the DASC descriptor with symmetric and asymmetric weights. (from left to right) Average bad-pixel error rates optimized with WTA under illumination and exposure variations, and GC under illumination and exposure variations.

Table 1 reports the computational complexity of the DASC descriptor for the brute-force implementation and the proposed efficient implementation when using the symmetric and asymmetric weights. The DASC descriptor with asymmetric weights provides a low computational complexity thanks to its efficient computational framework.

image size	SIFT [13]	DAISY [17]	LSS [14]	DASC* w/ sym.	DASC†w/ sym.	DASC* w/ asym.	DASC†w/ asym.
463×370	$130.3s$	$2.5s$	$31s$	$197.2s$	$9s$	$128s$	$5s$

Table 1. Evaluation of the computational complexity. The brute-force and efficient implementation of the DASC is denoted as * and †, respectively. However, the DASC descriptor with symmetric weights need more computational load compared to that of asymmetric weights.

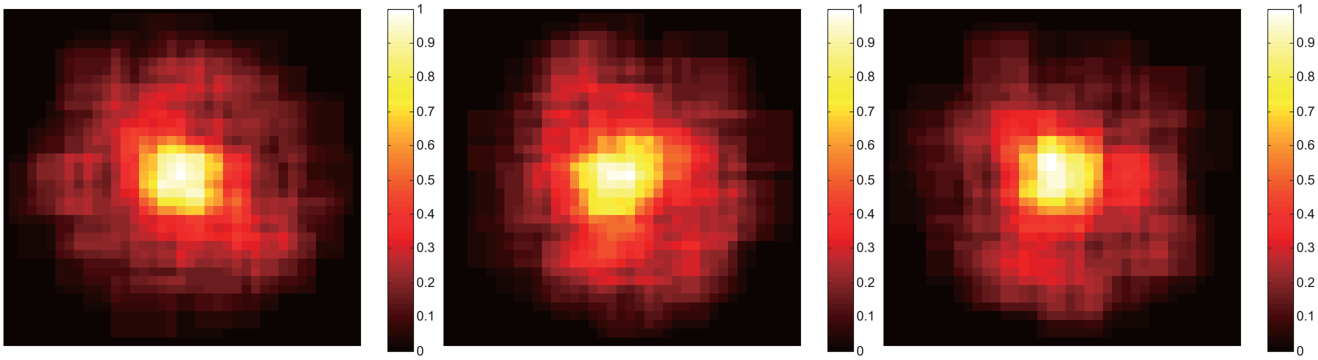
3. Multi-modal and Multi-spectral Feature Learning

In this section we provide an example of training pairs, denoted as $\mathcal{P} = \{(\mathcal{R}_m^1, \mathcal{R}_m^2, y_m)|m = 1, ..., N_t\}$, used in the sampling pattern learning where $(\mathcal{R}^1, \mathcal{R}^2)$ are support window pairs, and N_t 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 from ground truth correspondence maps for images captured under varying illumination conditions and/or with imaging devices [1, 5]. It should be noted that since multi-modal and multi-spectral pairs do not have a ground truth dense correspondence, we manually obtained ground truth displacement vectors [16]. In our experiments, we first established 50,000 multi-spectral and multi-modal support window pairs, as shown in Fig. 4. Among them, 5,000 matching support window pairs (positive samples, *i.e.*, $y_m = 1$) were randomly selected from true matching pairs, while 5,000 non-matching support window pairs (negative samples, *i.e.*, $y_m = 0$) were made by randomly selecting two support windows from different matching pairs. Thus, in total, $N_t = 10,000$ training support window pairs were built. In experiments, each training set is mutually used to learn a sampling pattern. Specifically, the sampling pattern for Middlebury benchmark data set is learned from the multi-spectral and multi-modal benchmark. In a similar way, the sampling patterns for multi-modal and multi-spectral benchmark and MPI SINTEL benchmark are learned from MPI SINTEL benchmark and multi-modal and multi-spectral benchmark, respectively.

(a) Middlebury benchmark (b) Multi-spectral and Multi-modal (c) MPI SINTEL benchmark

Figure 4. Some examples of 50,000 support window training pairs built from Middlebury stereo benchmark, multi-spectral and multi-modal benchmark, and MPI optical flow benchmark.

Fig. 5 shows patch-wise receptive fields on learned sampling patterns used in our DASC descriptor. For an effective visualization, we followed the practice used in [8]. We stacked all patch-wise receptive fields learnt from the Middlebury stereo benchmark [1], the multi-modal and multi-spectral benchmark [16, 3, 15, 9, 12], and the MPI SINTEL benchmark [5], respectively. A set of histogram bins corresponding to the patch of each patch-wise receptive field are incremented by one, and they are finally normalized with the maximum value. The density of patch-wise receptive fields tends to be concentrated on the center. In many literature, it has been shown that such a center-biased density distribution pooling in the local feature provides the robustness [2, 8].



(a) Middlebury benchmark

(b) Multi-spectral and Multi-modal

(c) MPI SINTEL benchmark

Figure 5. Visualization of patch-wise receptive fields of the DASC descriptor which are learned from Middlebury benchmark, multi-spectral and multi-modal benchmark, and MPI SINTEL benchmark.

4. More Results

In this section, we fist analyze the effects of the support window size and the feature dimension in our DASC descriptor. Then, we provide the additional results for our DASC descriptor and state-of-the-art descriptor-based methods and area-based methods using the Middlebury stereo benchmark, the multi-modal and multi-spectral image pair benchmark, and the MPI SINTEL optical flow benchmark.

4.1. Effects of Window Size and Feature Dimension

We analyze the effects of the support window size M and a feature dimension L in our DASC descriptor on the Middlebury stereo benchmark. We evaluate the performance by varying M from 5×5 to 33×33 and L from 50 to 400, respectively. It is worth noting that the computational complexity of our DASC descriptor is independent of the support window size M, since we extract the sampling patterns through the receptive field pooling from the support window. Instead, its complexity linearly increases in proportional to the number of sampling patterns, *i.e.*, the feature dimension L.

Fig. 6 shows the stereo matching results obtained with varying M. As expected, using small support windows degenerates the matching quality. In our paper, we used 31×31 as the support window size. Fig. 7 shows that the accuracy of the DASC descriptor is saturated when L is between $150 \sim 200$. Considering the trade-off between the accuracy and the runtime efficiency, we set the feature dimension L to 128.

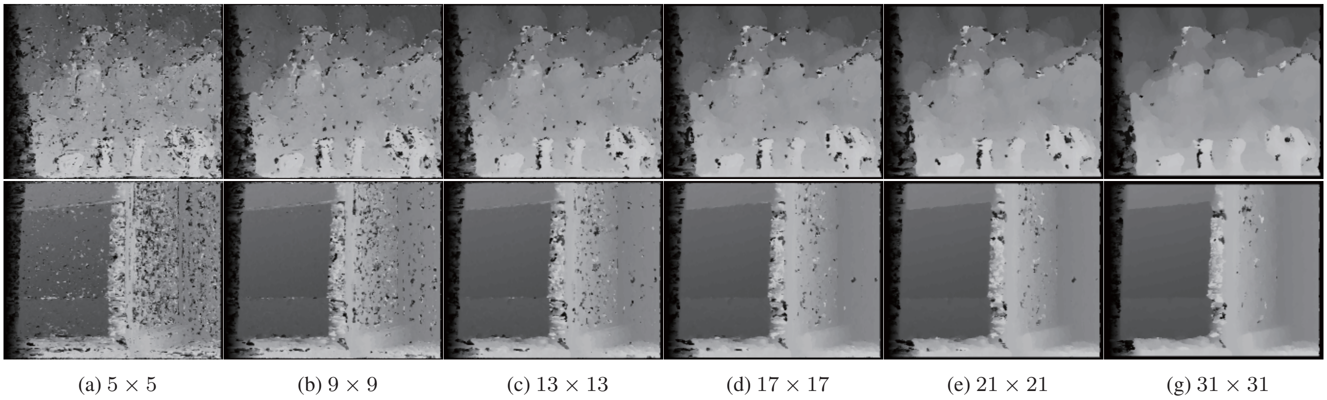

(a) 5×5 (b) 9×9 (c) 13×13 (d) 17×17 (e) 21×21 (g) 31×31

Figure 6. Results of disparity estimation for *Dolls* and *Wood1* image pairs taken under exposure combination '0/1' by varying the support window size M in the DASC descriptor. In our work, we used $N = 31 \times 31$ as the size of support window.

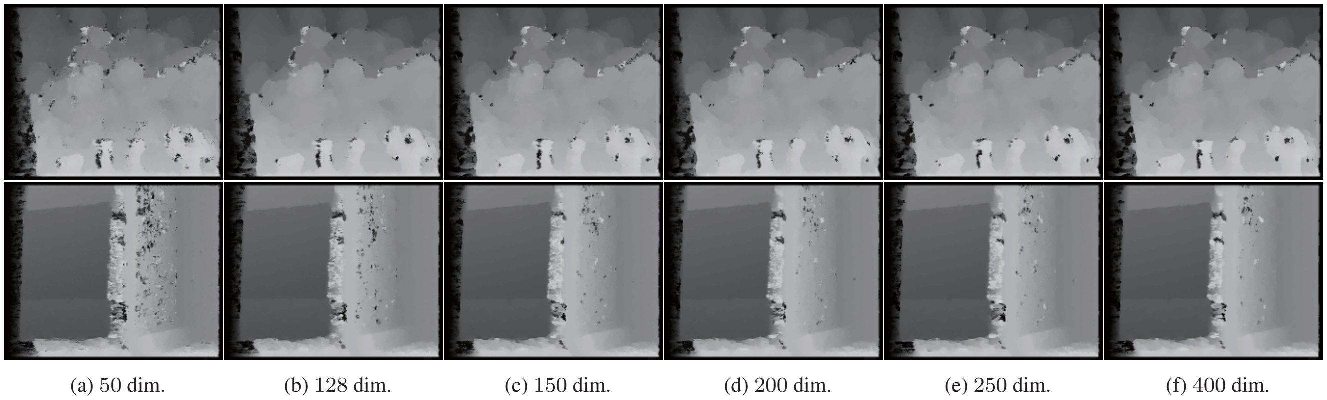

(a) 50 dim. (b) 128 dim. (c) 150 dim. (d) 200 dim. (e) 250 dim. (f) 400 dim.

Figure 7. Results of disparity estimation for *Dolls* and *Wood1* image pairs taken under exposure combination '0/1' by varying the descriptor dimension L in the DASC descriptor. In our work, we used $L = 128$ as the length of descriptor dimension.

4.2. Middlebury Stereo Benchmark

In Middlebury stereo benchmark, we used the *Art*, ***Baby1***, ***Books***, *Bowling2*, *Cloth3*, *Cloth4*, ***Dolls***, ***Moebius***, ***Reindeer***, and *Wood1*. In this supplementary materials, the results for **bold** image pairs are shown. Fig. 8, 9, 10, 11, and 12 compare the disparity maps estimated for stereo image pairs taken with an exposure combination '0/2'.

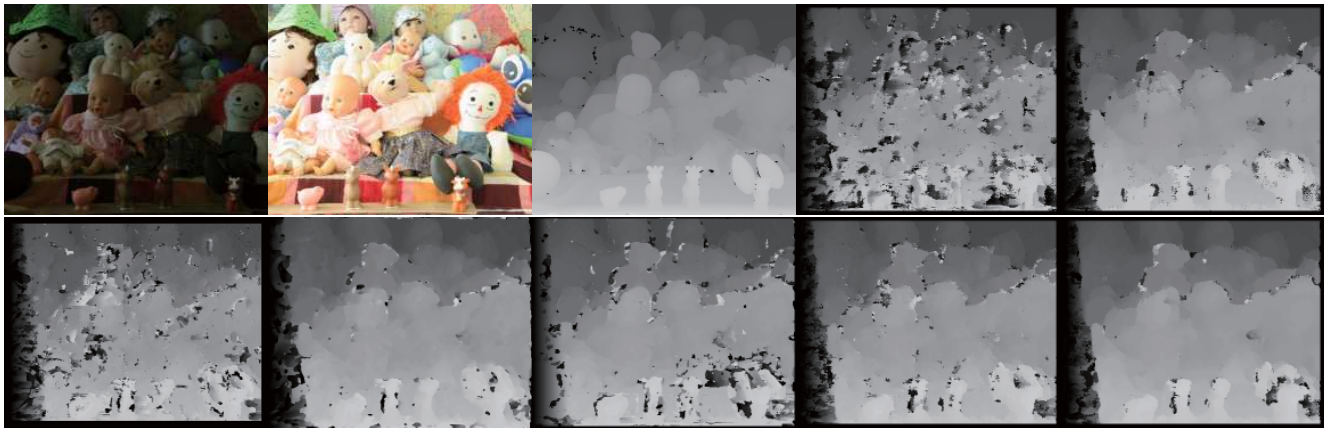


Figure 8. Comparison of disparity estimation for *Dolls* image pairs taken under illumination combination '0/2'. (from left to right, top and bottom) Left color image, right color image, and disparity maps for the ground truth, ANCC [11], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

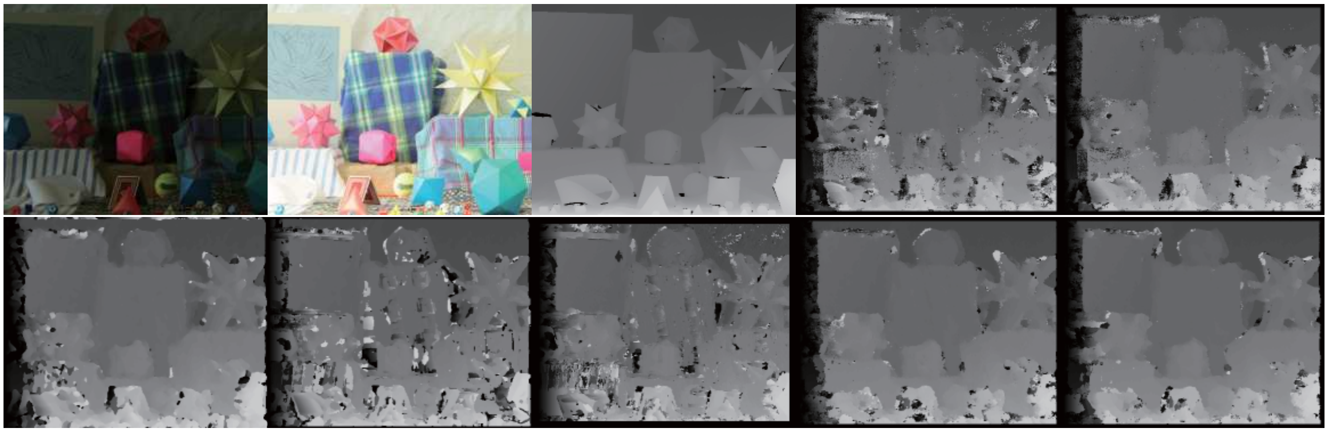


Figure 9. Comparison of disparity estimation for *Moebius* image pairs taken under illumination combination '0/2'. (from left to right, top and bottom) Left color image, right color image, and disparity maps for the ground truth, ANCC [11], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

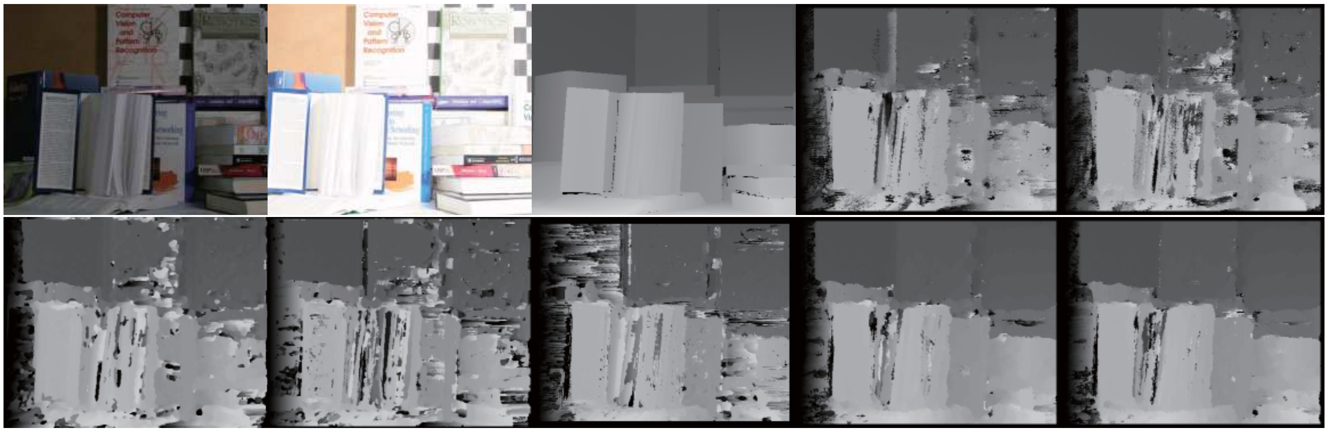


Figure 10. Comparison of disparity estimation for *Books* image pairs taken under illumination combination '0/2'. (from left to right, top and bottom) Left color image, right color image, and disparity maps for the ground truth, ANCC [11], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

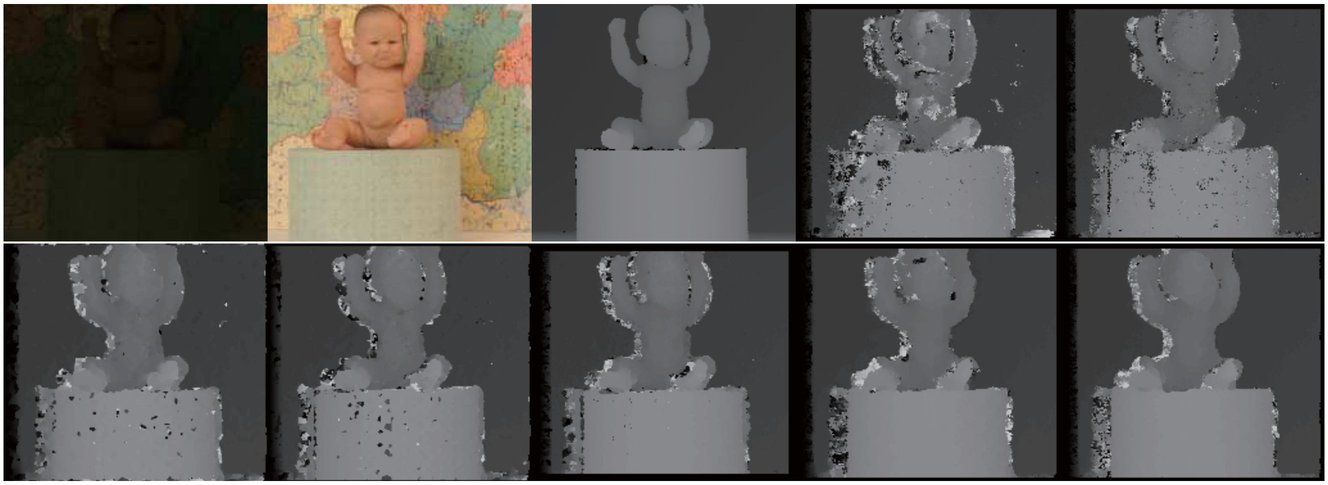


Figure 11. Comparison of disparity estimation for *Baby1* image pairs taken under illumination combination '0/2'. (from left to right, top and bottom) Left color image, right color image, and disparity maps for the ground truth, ANCC [11], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

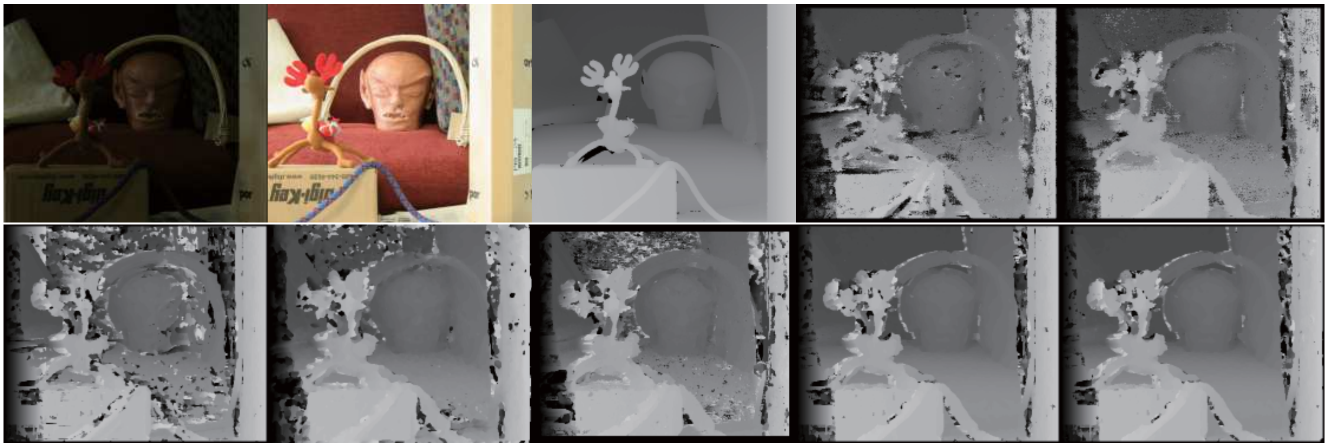


Figure 12. Comparison of disparity estimation for *Reindeer* image pairs taken under illumination combination '0/2'. (from left to right, top and bottom) Left color image, right color image, and disparity maps for the ground truth, ANCC [11], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

4.3. Multi-modal and Multi-spectral Image Pairs

In experiments, the multi-modal and multi-spectral image pairs consist of RGB-NIR images, flash-noflash images, images taken under different exposures, and blurred-clean images.

- RGB-NIR image pairs: *epfl1*, *epfl2*, *epfl3*, *epfl4*, *epfl5*, ***epfl6***, ***lion***, *myrgbnir*, ***orchid***, *stereo3*, and *stereo4*.
- Flash-noflash image pairs: ***Dolls1***, ***Dolls2***, and ***Dolls3***.
- Image pairs taken under different exposures: *altar*, *BabyAtWindow*, *BabyOnGrass*, ***balcony***, *books*, *ChristmasRider*, *clouds*, *FeedingTime*, *flower*, *HighChair*, *LadyEating*, ***lantern***, *mpi*, *PianoMan*, ***room***, *SantasLittleHelper*, *street*, and *window*.
- Blurred-clean image pairs: *avisar*, ***books1***, *books2*, ***cars1***, *cars2*, *children*,***face1***, *face2*, *flowers*, , *numbers*, and *yemin*.

In this supplementary materials, the results for **bold** image pairs are shown. Fig. 13, 14, 15, and 16 show the warped color image and its corresponding 2-D flow fields for multi-modal and multi-spectral image pairs. For the results of objective comparison, please refer to Table 2 in our paper.

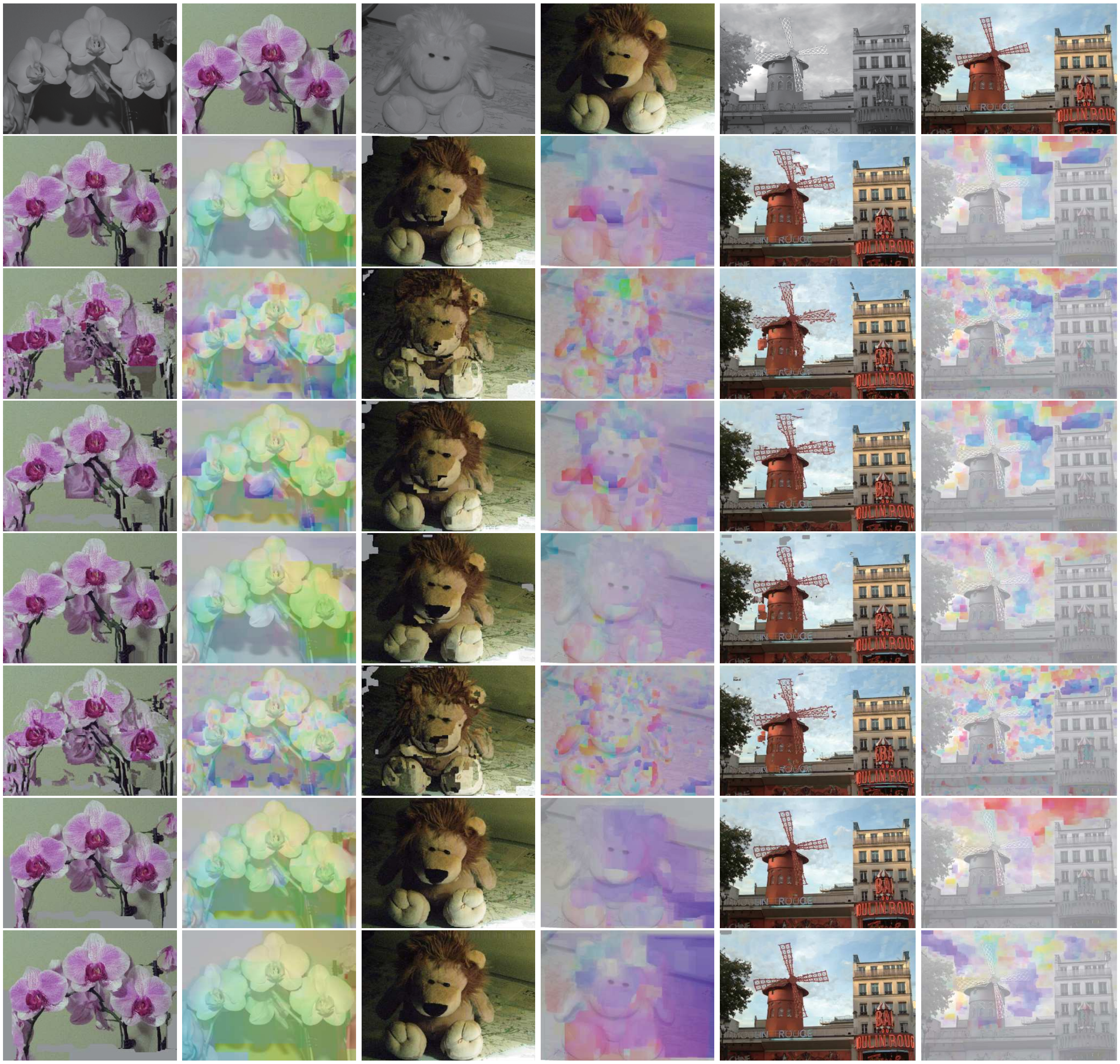


Figure 13. Comparison of dense correspondence for RGB-NIR images including *orchid*, *lion*, and *epfl6*. (from top to bottom) Input image pairs, RSNCC [16], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

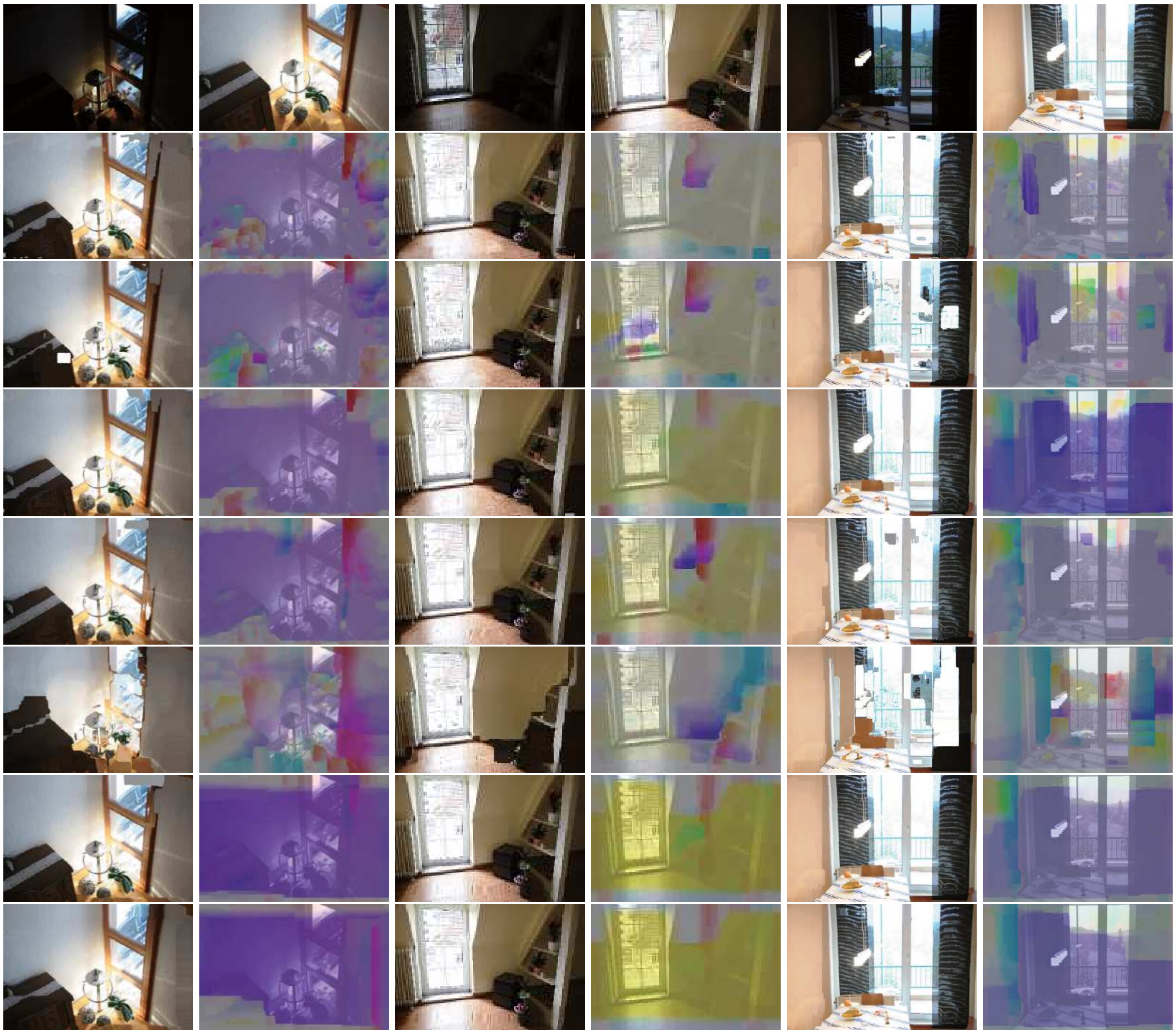


Figure 14. Comparison of dense correspondence for different exposure images including *lantern*, *balcony*, and *room* . (from top to bottom) Input image pairs, RSNCC [16], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

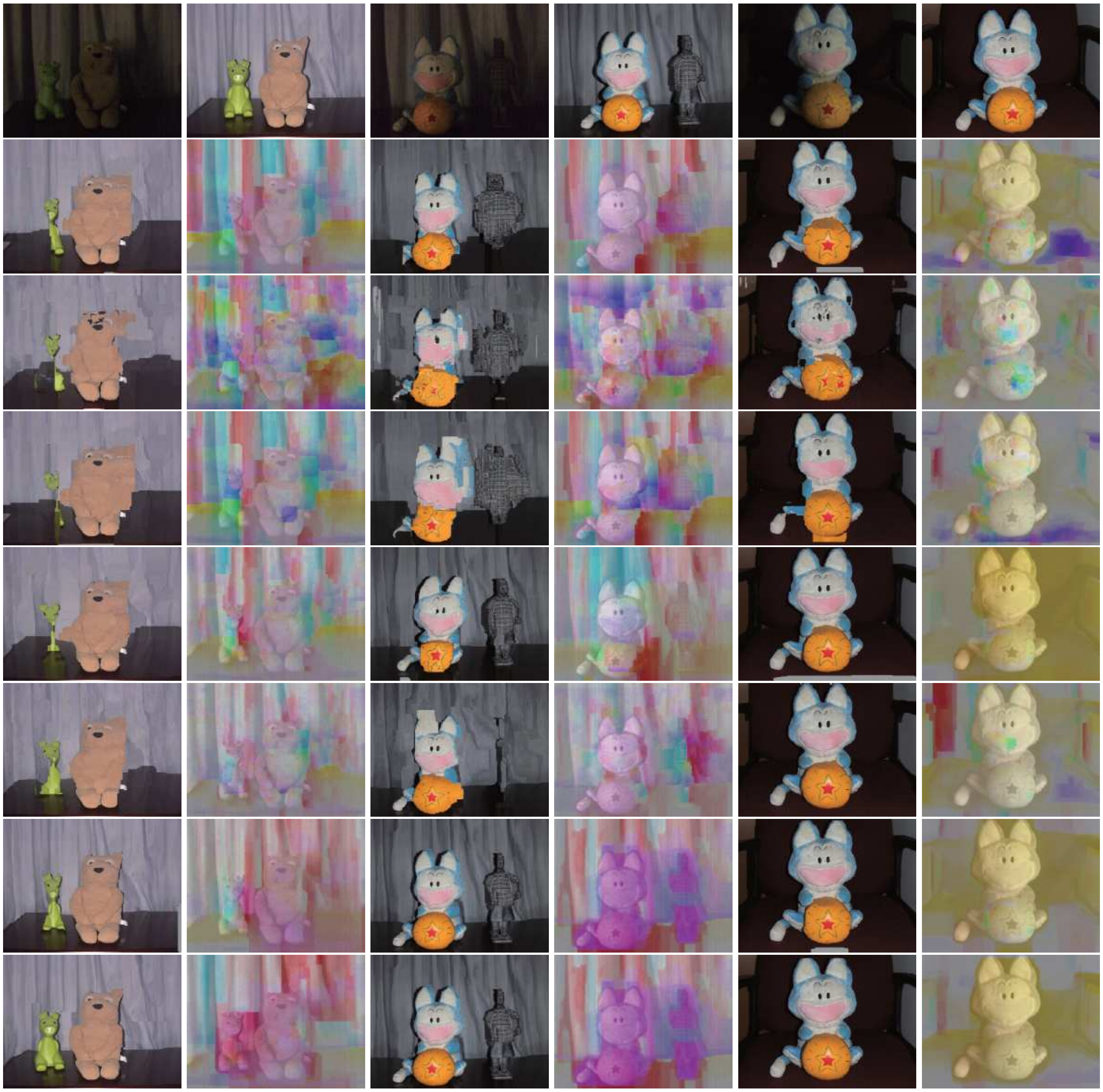


Figure 15. Comparison of dense correspondence for flash-noflash images including *Dolls1*, *Dolls2*, and *Dolls3*. (from top to bottom) Input image pairs, RSNCC [16], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

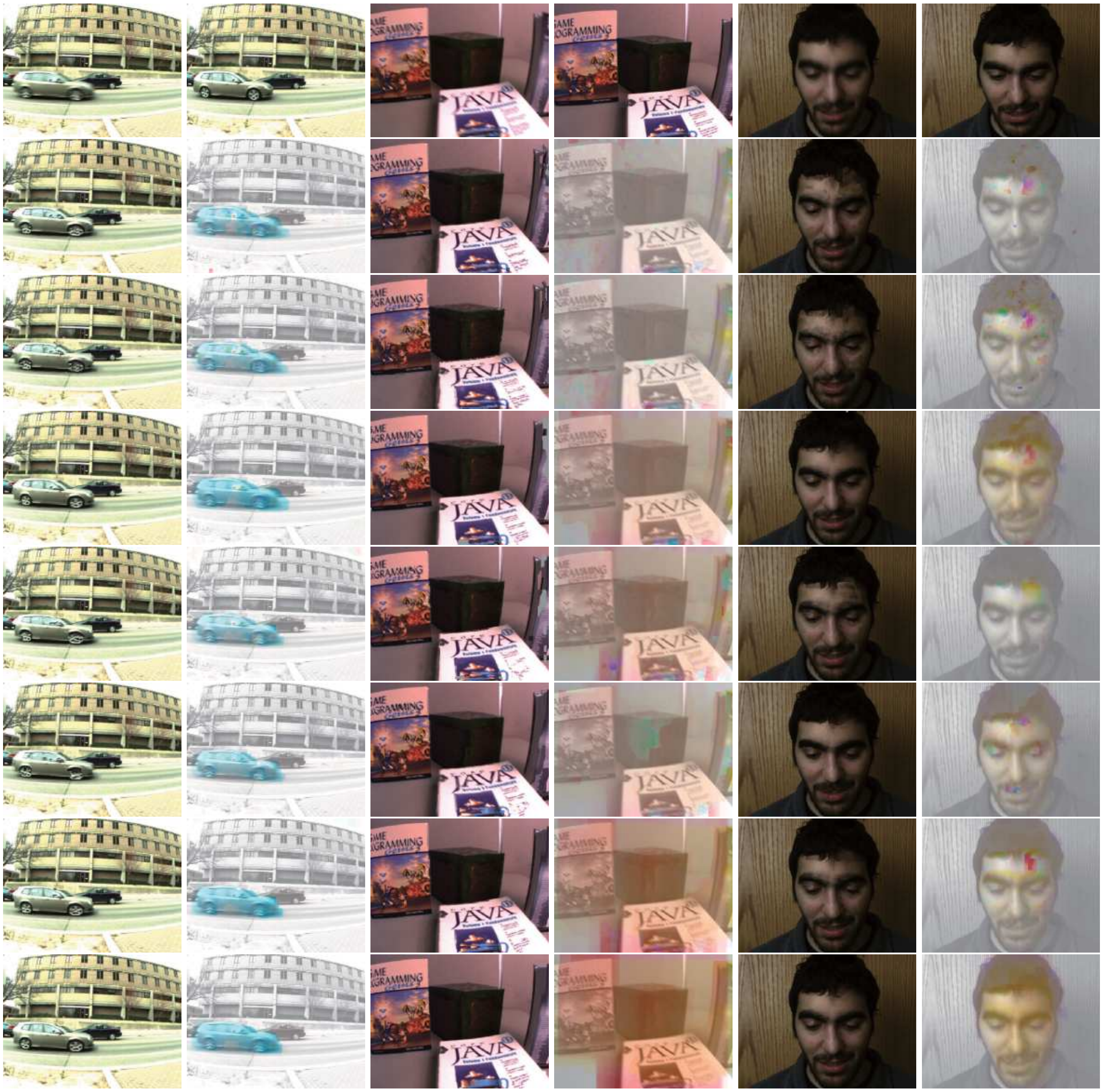


Figure 16. Comparison of dense correspondence for blurred images *cars1*, *books1*, and *face1*. (from top to bottom) Input image pairs, RSNCC [16], BRIEF [6], DAISY [17], SIFT [13], LSS [14], DASC+RP, and DASC+LRP.

4.4. MPI SINTEL Optical Flow Benchmark

In MPI SINTEL optical flow benchmark, the dataset consists of two kind of rendering frames, namely *clean pass* and *final pass*, each containing 12 sequences with over 500 frames in total [5]. Fig. 17 shows visual comparison on the MPI SINTEL benchmark, where the warped color image and its corresponding 2-D flow fields are depicted.

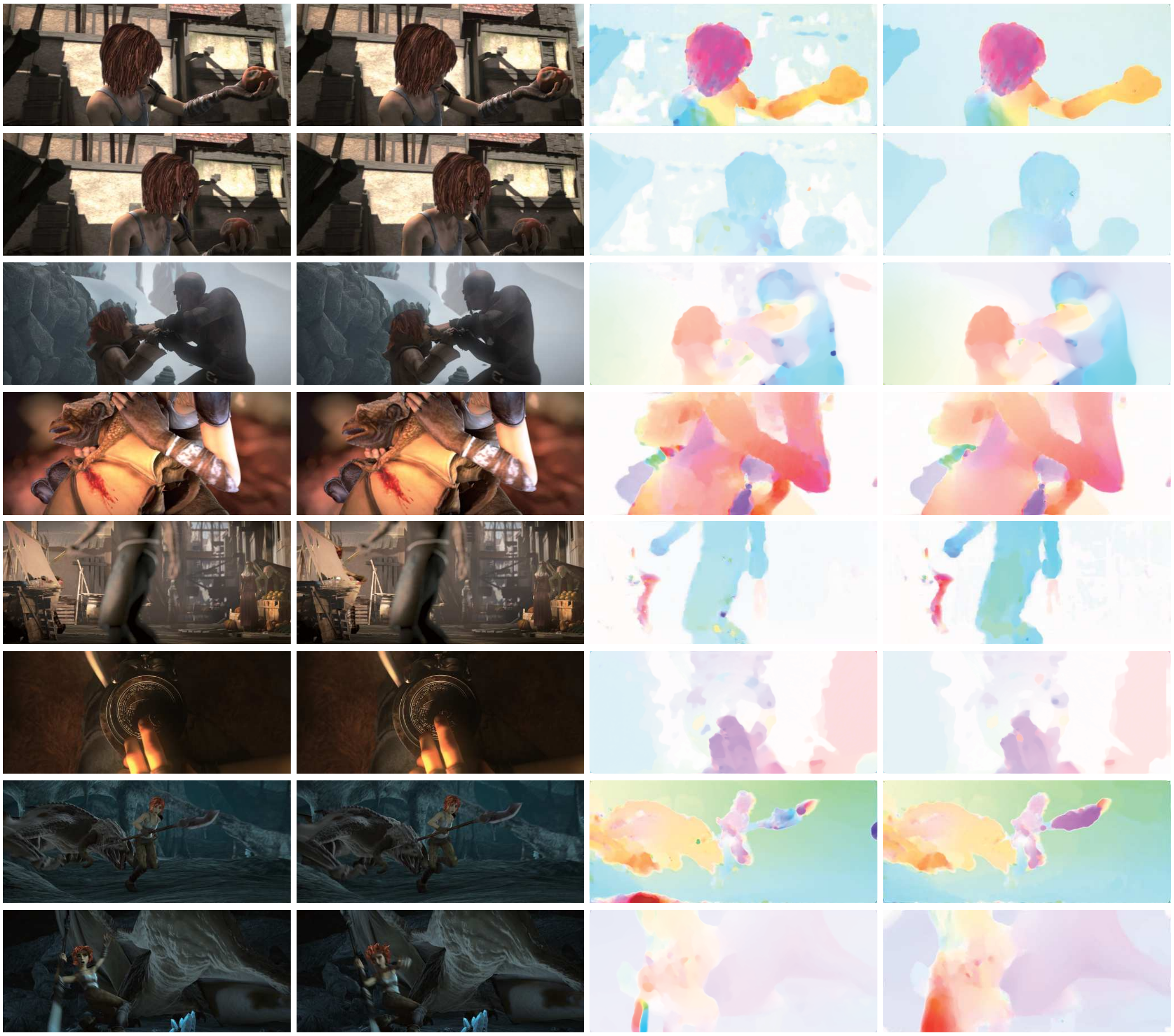


Figure 17. Visual comparison on the MPI Sintel benchmark. (from left to right) Input image 1 and 2, flow field estimation results of LDOF [4] and LDOF with the DASC+LRP descriptor. Note that the histogram of oriented gradient (HOG) [7] is used in the original LDOF [4].

References

[1] http://vision.middlebury.edu/stereo/.

[2] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak : Fast retina keypoint. *In Proc. of CVPR*, 2012.

[3] M. Brown and S. Susstrunk. Multispectral sift for scene category recognition. *In Proc. of CVPR*, 2011.

[4] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. PAMI*, 33(3):500–513, 2011.

[5] D. Butler, J. Wulff, G. Stanley, and M. Black. A naturalistic open source movie for optical flow evaluation. *In Proc. of ECCV*, 2012.

[6] M. Calonder. Brief : Computing a local binary descriptor very fast. *IEEE Trans. PAMI*, 34(7):1281–1298, 2011.

[7] N. Dalal and B. Trigg. Histograms of oriented gradients for human detection. *In Proc. of CVPR*, 2005.

[8] B. Fan, Q. Kong, T. Trzcinski, and Z. Wang. Receptive fields selection for binary feature description. *IEEE Trans. IP*, 23(6):2583–2595, 2014.

[9] Y. HaCohen, E. Shechtman, and E. Lishchinski. Deblurring by example using dense correspondence. *In Proc. of ICCV*, 2013.

[10] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE Trans. PAMI*, 35(6):1397–1409, 2013.

[11] Y. Heo, K. Lee, and S. Lee. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Trans. PAMI*, 33(4):807–822, 2011.

[12] H. Lee and K. Lee. Dense 3d reconstruction from severely blurred images using a single moving camera. *In Proc. of CVPR*, 2013.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[14] E. Schechtman and M. Irani. Matching local self-similarities across images and videos. *In Proc. of CVPR*, 2007.

[15] 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 SIGGRAGH*, 2012.

[16] X. Shen, L. Xu, Q. Zhang, and J. Jia. Multi-modal and multi-spectral registration for natural images. *In Proc. of ECCV*, 2014.

[17] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. PAMI*, 32(5):815–830, 2010.