Learning to Compare Image Patches via Convolutional Neural Networks

Sergey Zagoruyko and Nikos Komodakis

Ecole des Ponts ParisTech, Universite Paris-Est, France

The document includes additional experimental results. It is split in 3 main sections, related to evaluation on: local image patches benchmark [1], wide baseline stereo, and local descriptors benchmark [2].

1. Local image patches benchmark

1.1. l_2-decision networks

We provide here a more detailed quantitative comparison of l_2-decision networks (i.e., where we use l_2 distance to compare descriptors at test time). To that end, we show the corresponding ROC curves in figure 1, comparing also with the state-of-the-art method [3]. As can be observed, the $\mathrm{siam\text{-}2stream\text{-}}l_2$ model exhibits the best performance on all datasets combinations except when being tested on Yosemite.

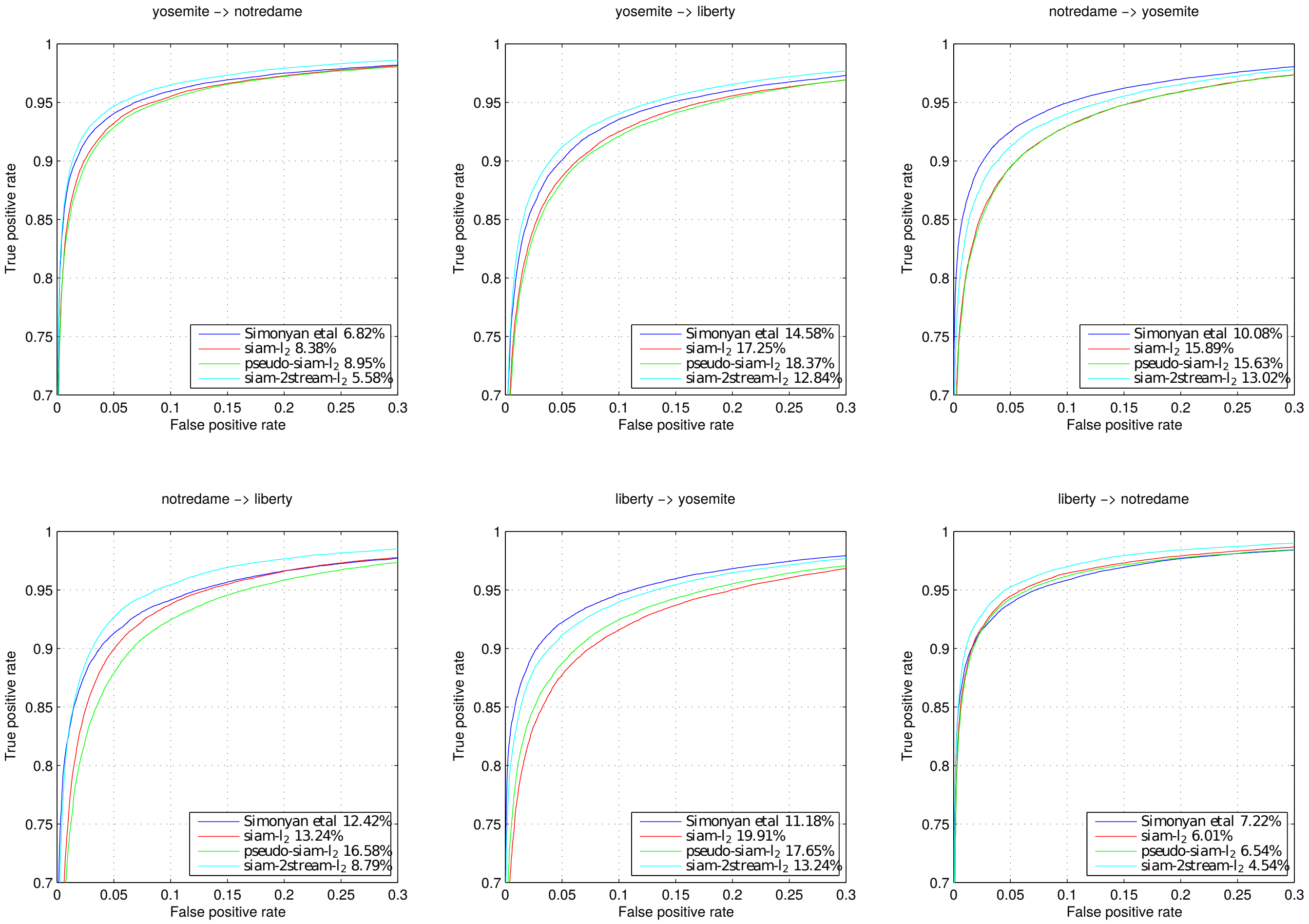


Figure 1. ROC curves of l_2 networks. $\mathrm{siam\text{-}2stream\text{-}}l_2$ shows the best performance on 4 out of 6 combinations of sequences

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

1.2. pseudo-siam **network**

The pseudo-siam network has two uncoupled branches which make it asymmetric. It is possible to make its decision symmetric by taking the sum of decisions from both possible combinations of patches in pair. Let P_1 and P_2 be the patches in pair and $o(P_1, P_2)$ - network's decision on these patches. Then the symmetric decision is defined as:

$$o_s(P_1, P_2) = o(P_1, P_2) + o(P_2, P_1) \tag{1}$$

In table 1 we show the results of evaluation of the above decision function. It's mean FPR95 over all dataset combinations is 9.11, which is by 0.63 better than a single asymmetric decision result and by 0.96 better than a result of siam network.

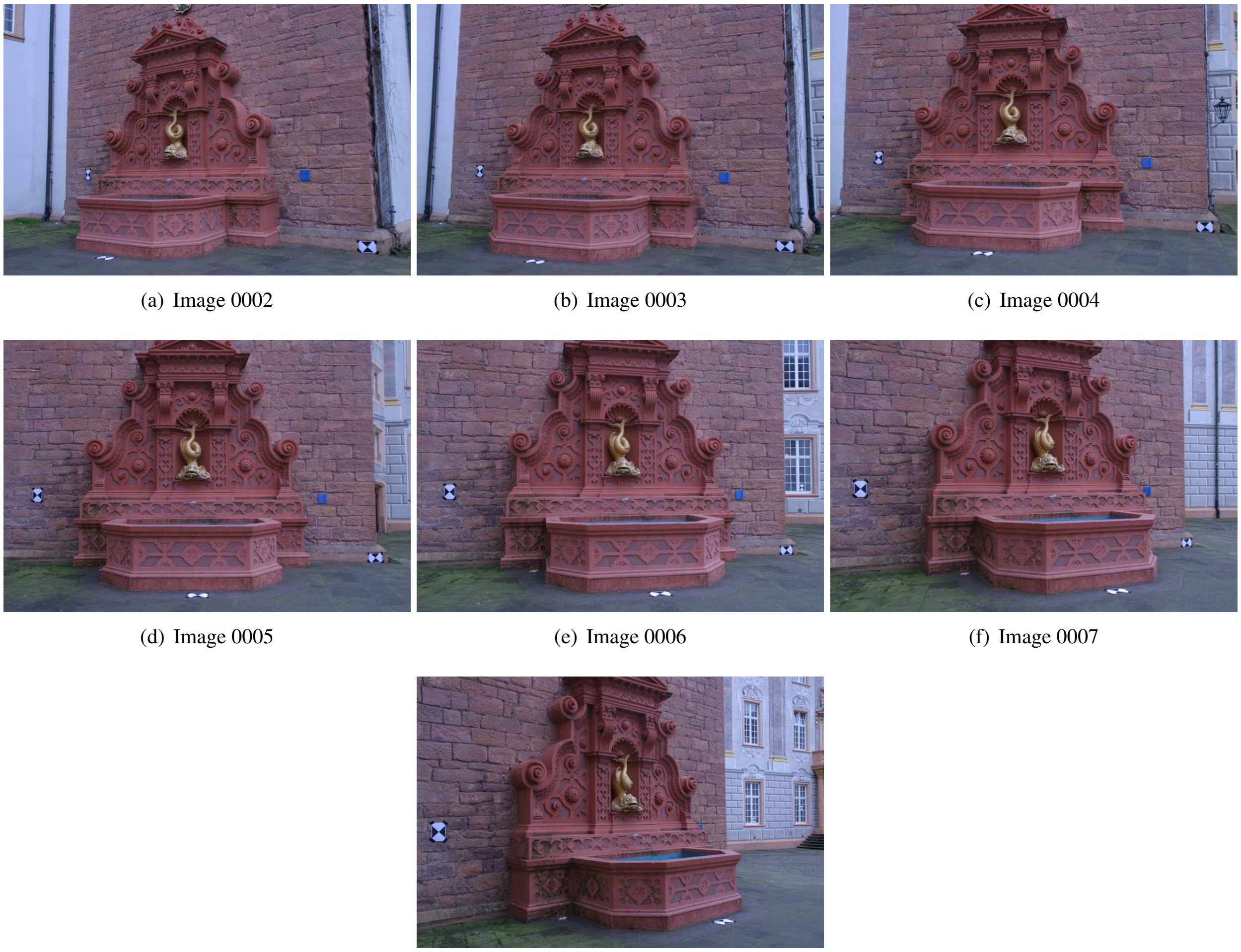
		$o(P_1, P_2)$	$o(P_1, P_2) + o(P_2, P_1)$
Yos	ND	5.44	4.82
Yos	Lib	12.64	11.79
ND	Yos	13.61	13.25
ND	Lib	10.35	9.99
Lib	Yos	12.50	11.44
Lib	ND	3.93	3.37
mean		9.74	9.11
mean(1,4)		10.51	9.96

Table 1. Results of pseudo-siam network with symmetric decision function evaluation

2. Wide baseline stereo evaluation

We show quantitative and qualitative evaluation results on “fountain” and “herzjesu” datasets from [4]. We compare our networks 2ch, siam-2stream-l_2, siam with the state of the art descriptor DAISY [5].

2.1. “Fountain” dataset



(a) Image 0002

(b) Image 0003

(c) Image 0004

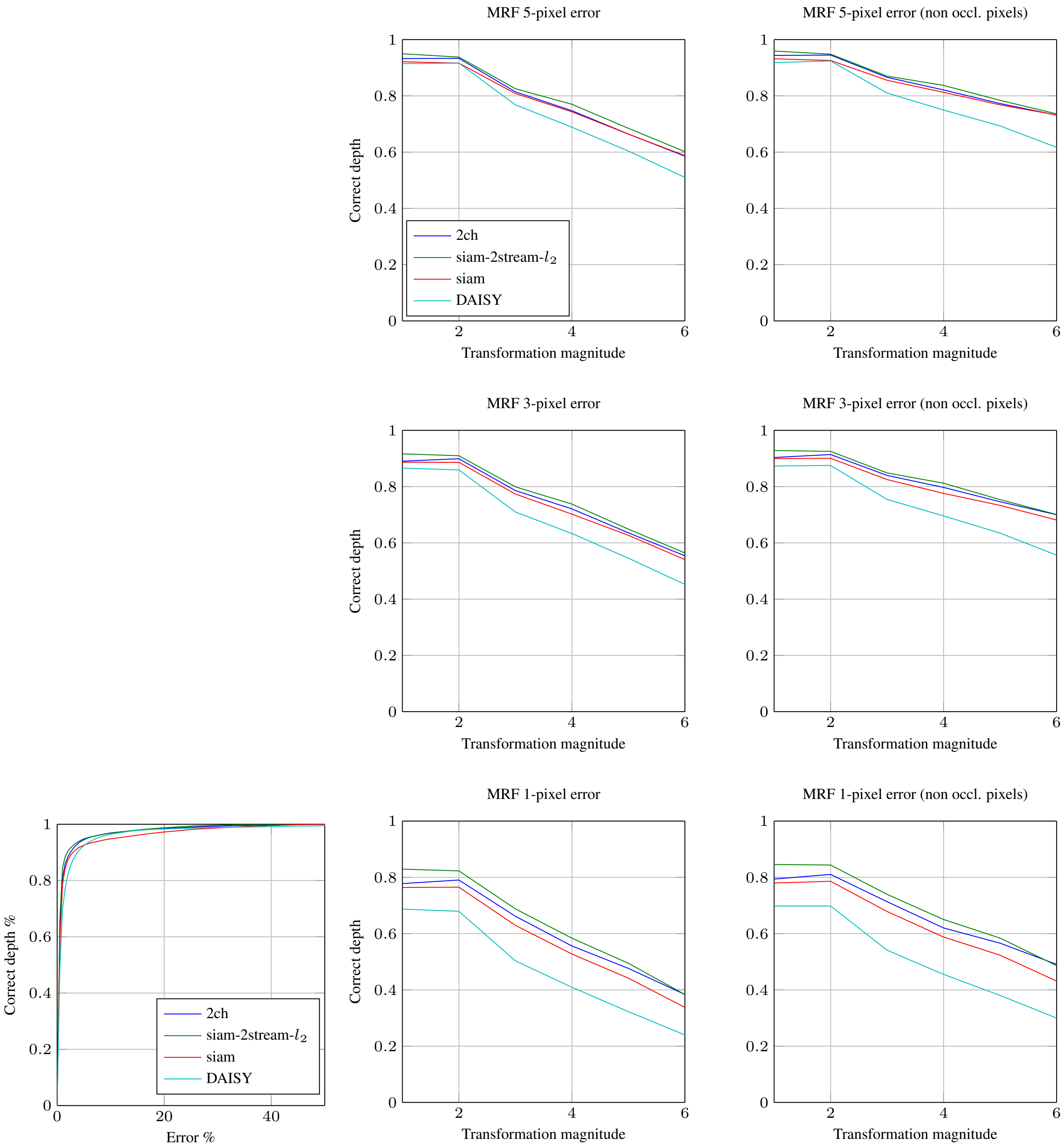
(d) Image 0005

(e) Image 0006

(f) Image 0007

(g) Image 0008

Figure 2. Images from ”fountain” dataset. We use images 0002-0008 to generate 6 rectified stereo pairs against image 0003



(a)

(b)

Figure 3. Quantitative comparison for wide baseline stereo evaluation on "fountain" dataset. (a) Distributions of deviations from the laser-scan data, expressed as a fraction of the scene's depth range of the second of the second depth map in the sequence. (b) Distribution of errors for stereo pairs of increasing baseline (horizontal axis) both with and without taking into account occluded pixels (error thresholds were set equal to 5, 3 and 1 pixels in these plots - maximum disparity is around 500 pixels).

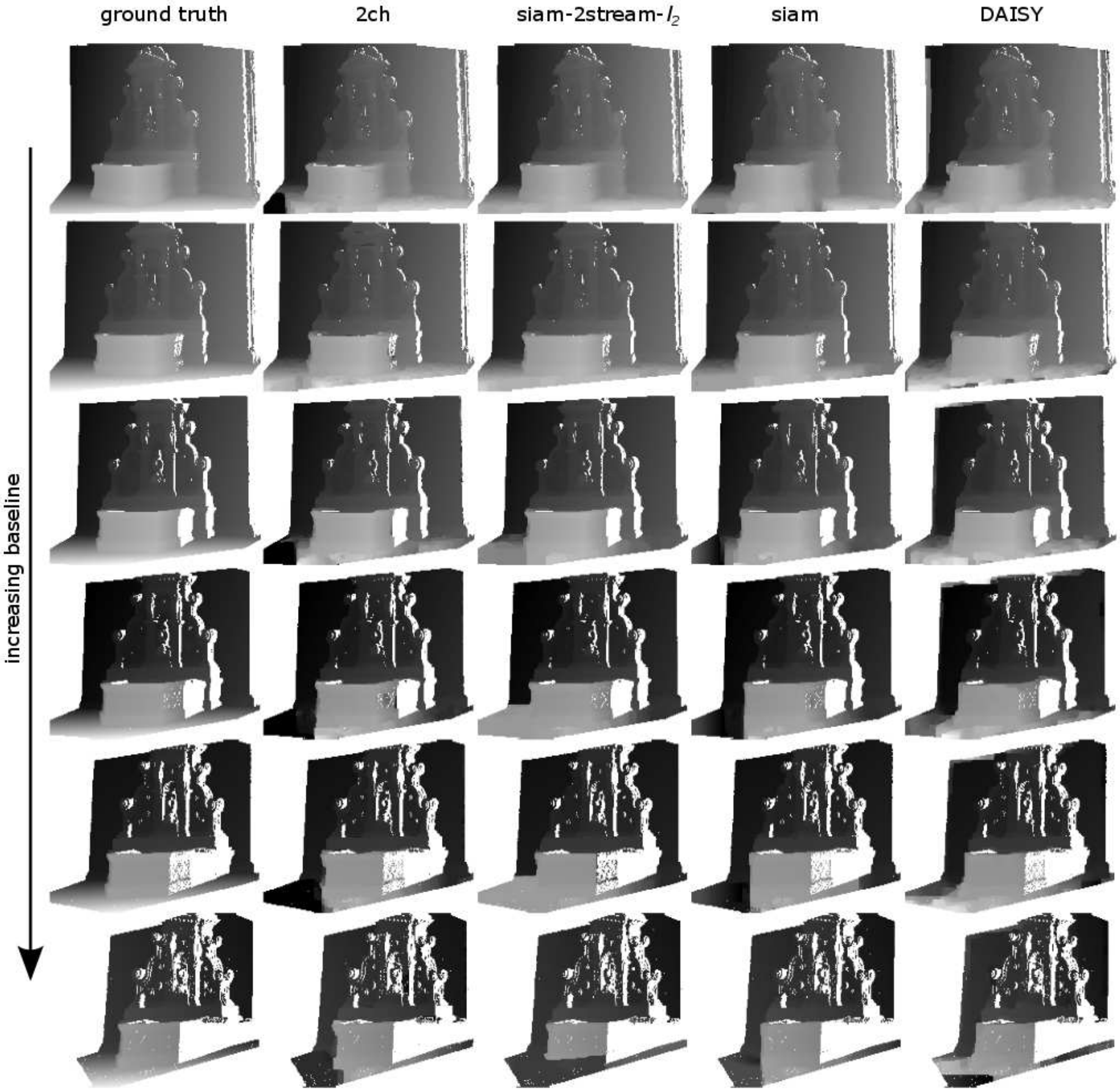
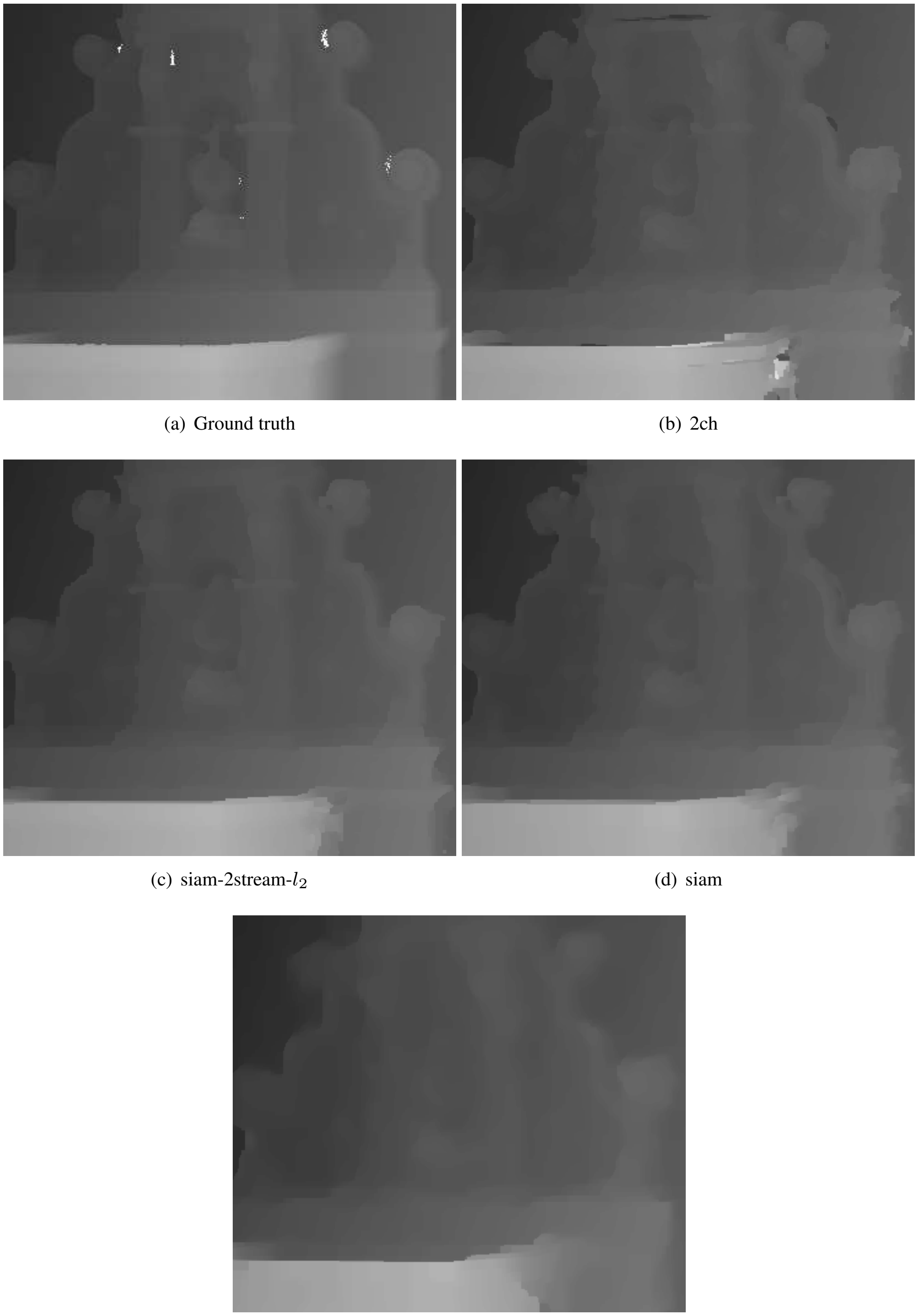


Figure 4. Qualitative comparison for wide baseline stereo evaluation on “fountain” dataset. From left to right column we show depth maps from ground truth, $\mathrm{2ch}$, $\mathrm{siam\text{-}2stream\text{-}}l_2$, siam networks and DAISY. The baseline between stereo pairs increases from top to bottom. All depth maps were computed with MRF optimization, only non-occluded pixels are shown.



(a) Ground truth

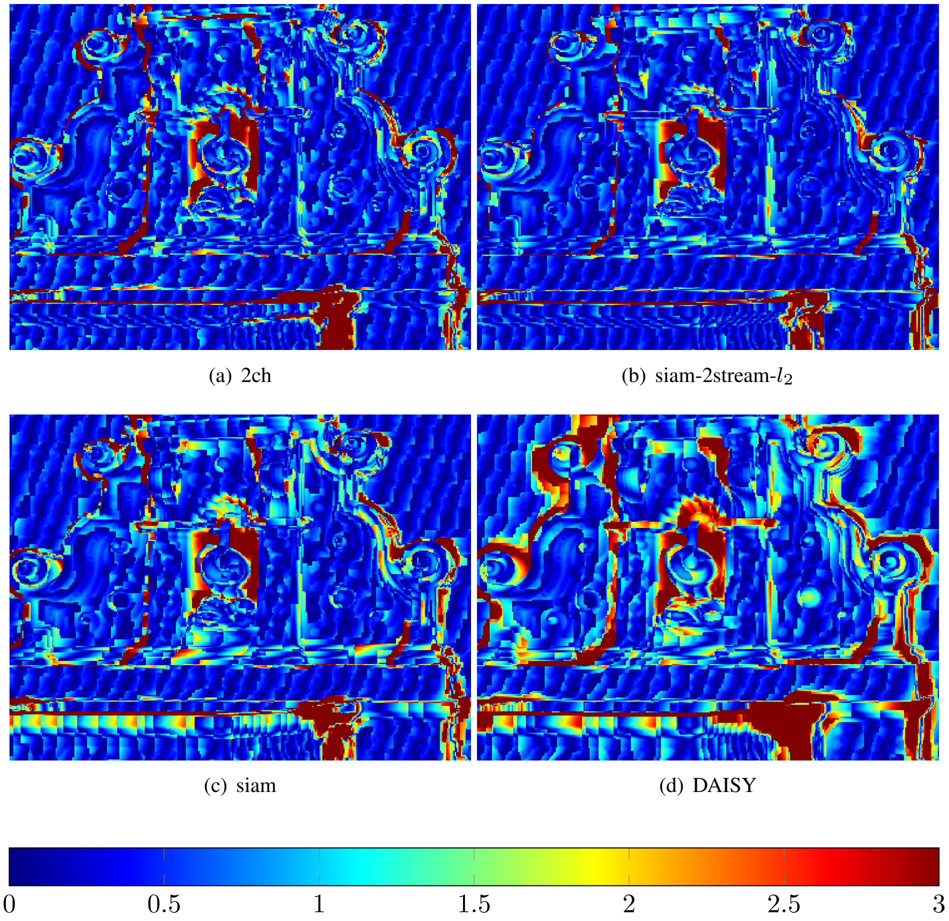
(b) 2ch

(c) siam-2stream-l_2

(d) siam

(e) DAISY

Figure 5. Close-up views on wide-baseline stereo evalutaion results on ”fountain” dataset.


(a) 2ch

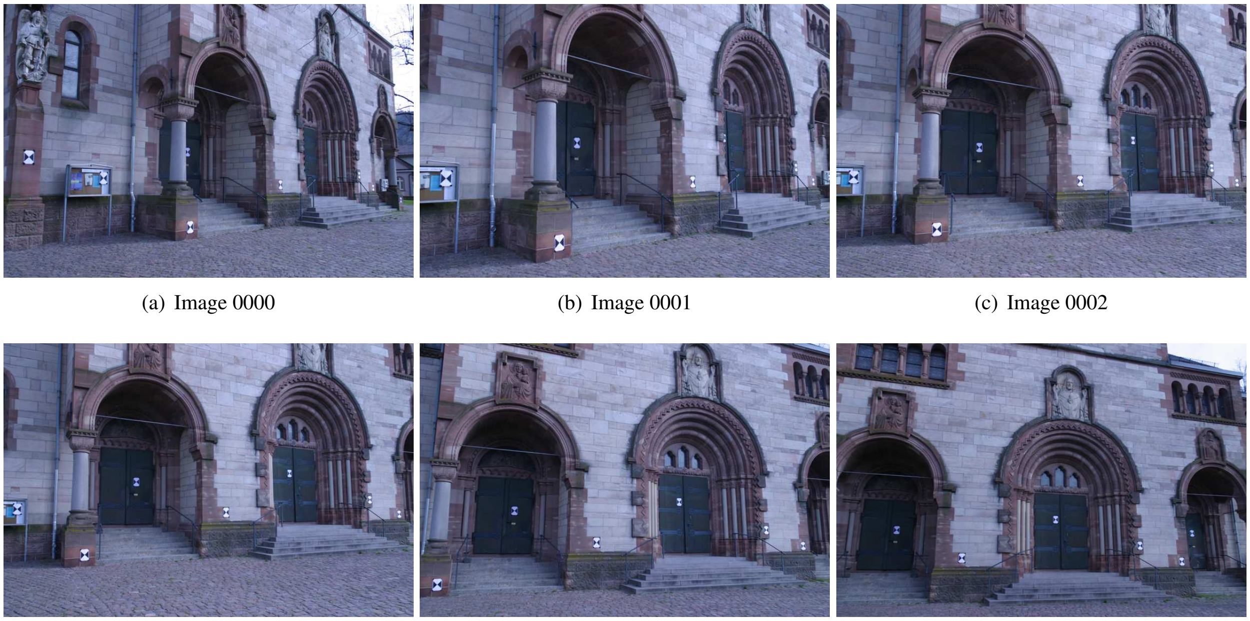
(b) siam-2stream-l_2

(c) siam

(d) DAISY

Figure 6. For the close-up views of fig. 5 we show thresholded absolute differences of ground truth depth map and estimated depth maps. Threshold is set to 3 pixels.

2.2. “Herzjesu” dataset



(a) Image 0000

(b) Image 0001

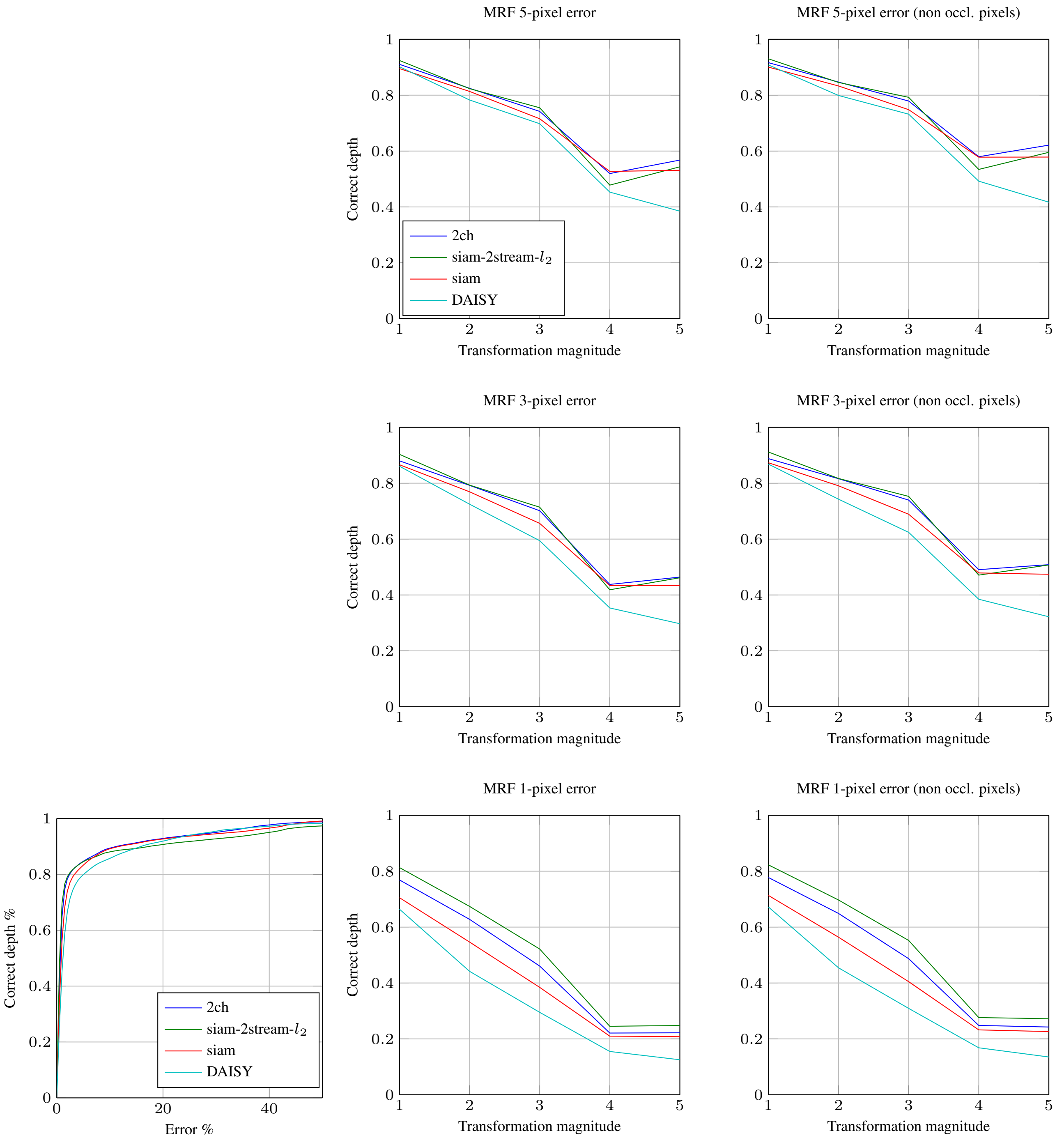
(c) Image 0002

(d) Image 0003

(e) Image 0004

(f) Image 0005

Figure 7. Images from ”herzjesu” dataset. We use images 0000-0005 to generate 5 stereo pairs against image 0005.



(a)

(b)

Figure 8. Quantitative comparison for wide baseline stereo on “herzjesu” dataset. (a) Distributions of deviations from the laser-scan data, expressed as a fraction of the scene’s depth range of the second of the second depth map in the sequence. (b) Distribution of errors for stereo pairs of increasing baseline (horizontal axis) both with and without taking into account occluded pixels (error thresholds were set equal to 5, 3 and 1 pixels in these plots - maximum disparity is around 500 pixels).

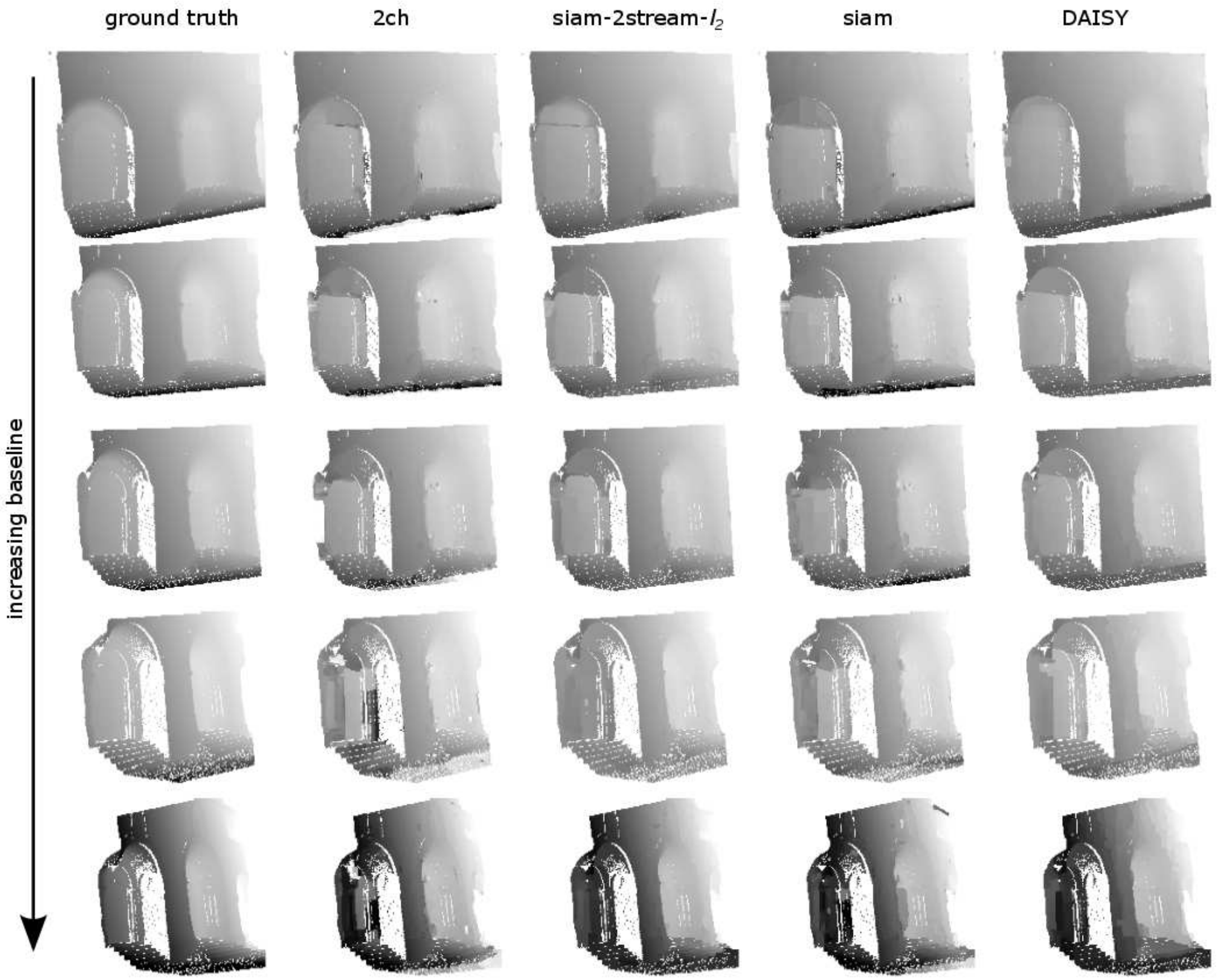
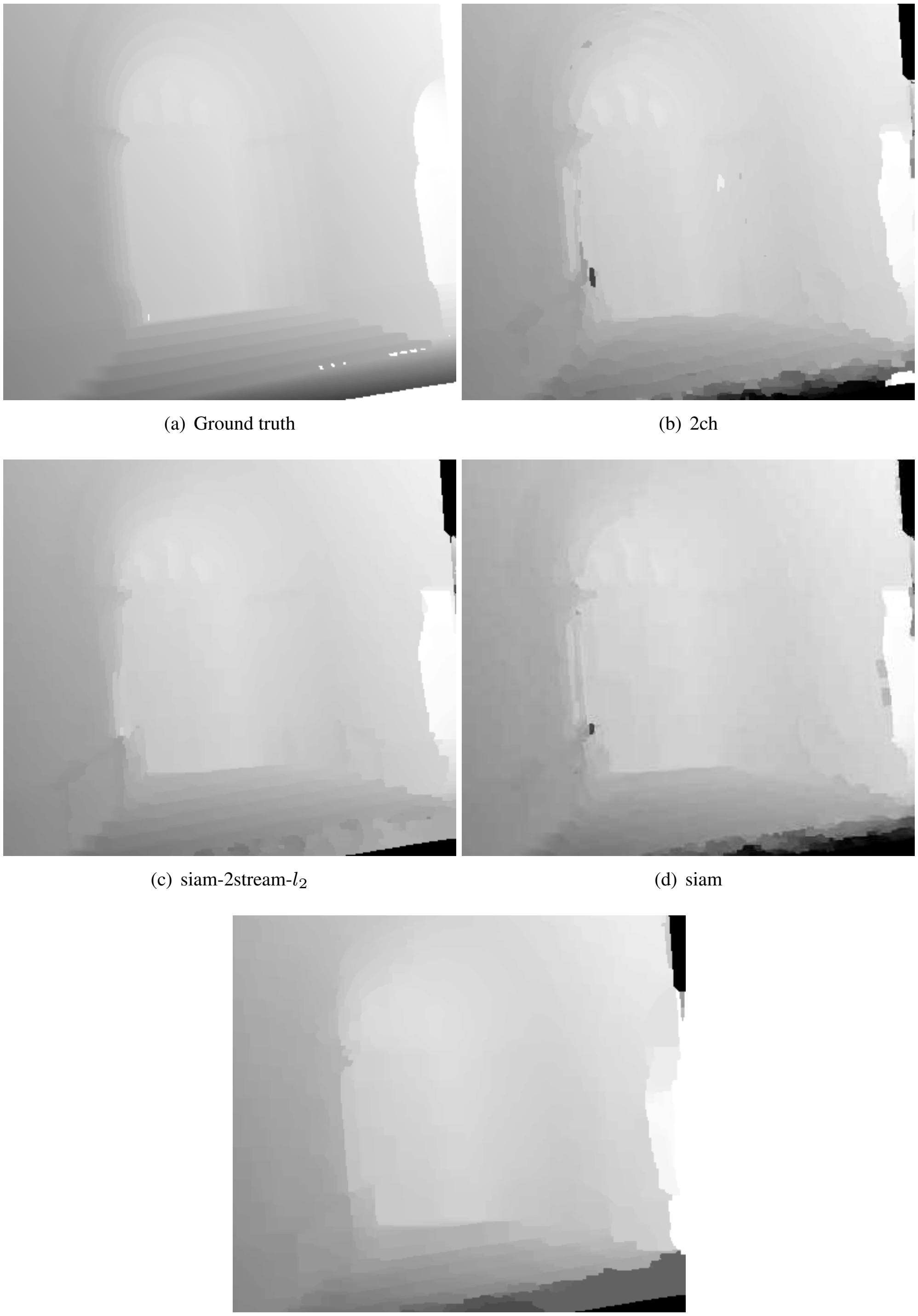


Figure 9. Qualitative comparison for wide baseline stereo evaluation on "herzjesu" dataset. From left to right column we show depth maps from ground truth, $\mathrm{2ch}$, $\mathrm{siam\text{-}2stream\text{-}}l_2$, siam networks and DAISY. The baseline between stereo pairs increases from top to bottom. All depth maps were computed with MRF optimization, only non-occluded pixels are shown.



(a) Ground truth

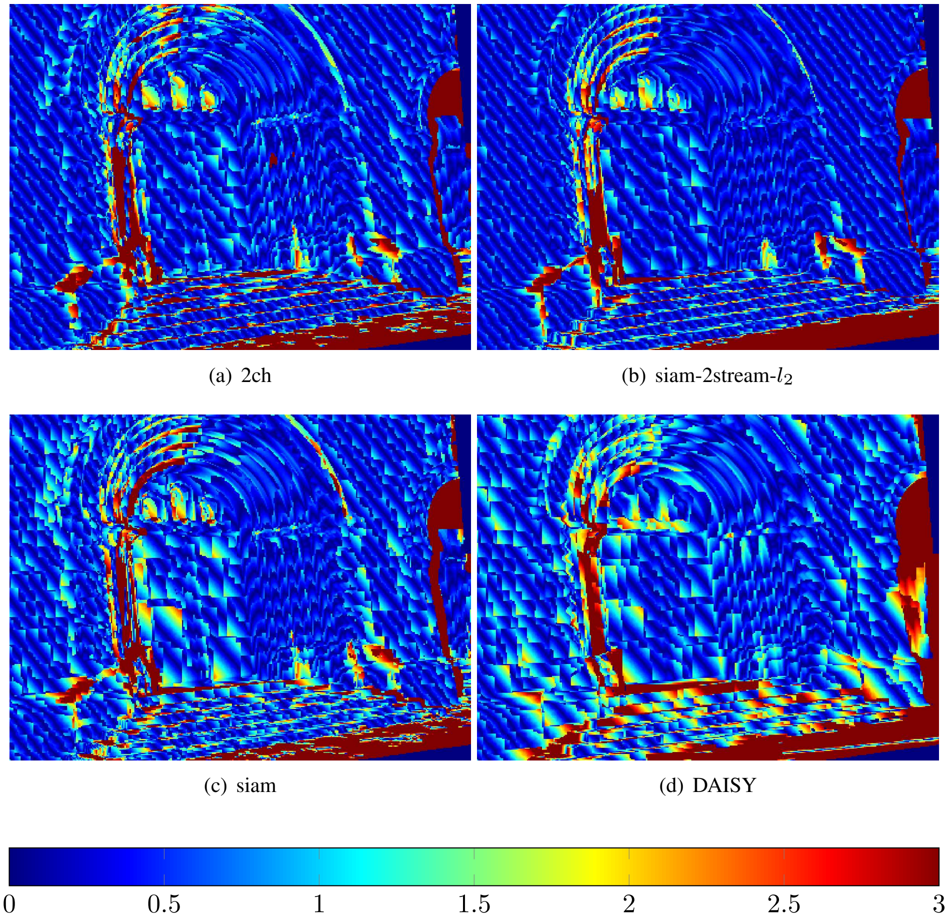
(b) 2ch

(c) siam-2stream-l_2

(d) siam

(e) DAISY

Figure 10. Close-up views on wide-baseline stereo evaluation results on ”herzjesu” dataset.


(a) 2ch

(b) siam-2stream-l_2

(c) siam

(d) DAISY

Figure 11. For the close-up views of fig. 10 we show thresholded absolute differences of ground truth depth map and estimated depth maps. Threshold is set to 3 pixels.

3. Local descriptors performance evaluation

We provide in fig. 12 evaluation plots for all sequences from Mikolajczyk dataset [2]. To compute the performance measure we extract elliptic regions of interest and corresponding image patches from both images using MSER detector. Minimal area size of detected ellipses set to 100. Next we compute the descriptors of all extracted patches and match all of them based on l_2 distance. A pair is a true positive if and only if the ellipse of the descriptor in the target image and the ground truth ellipse have an intersection over union that is greater than or equal to 0.6 (all other pairs are false positives). Based on this, a precision recall curve is computed and the area under this curve (average precision) is used as performance measure (mAP).

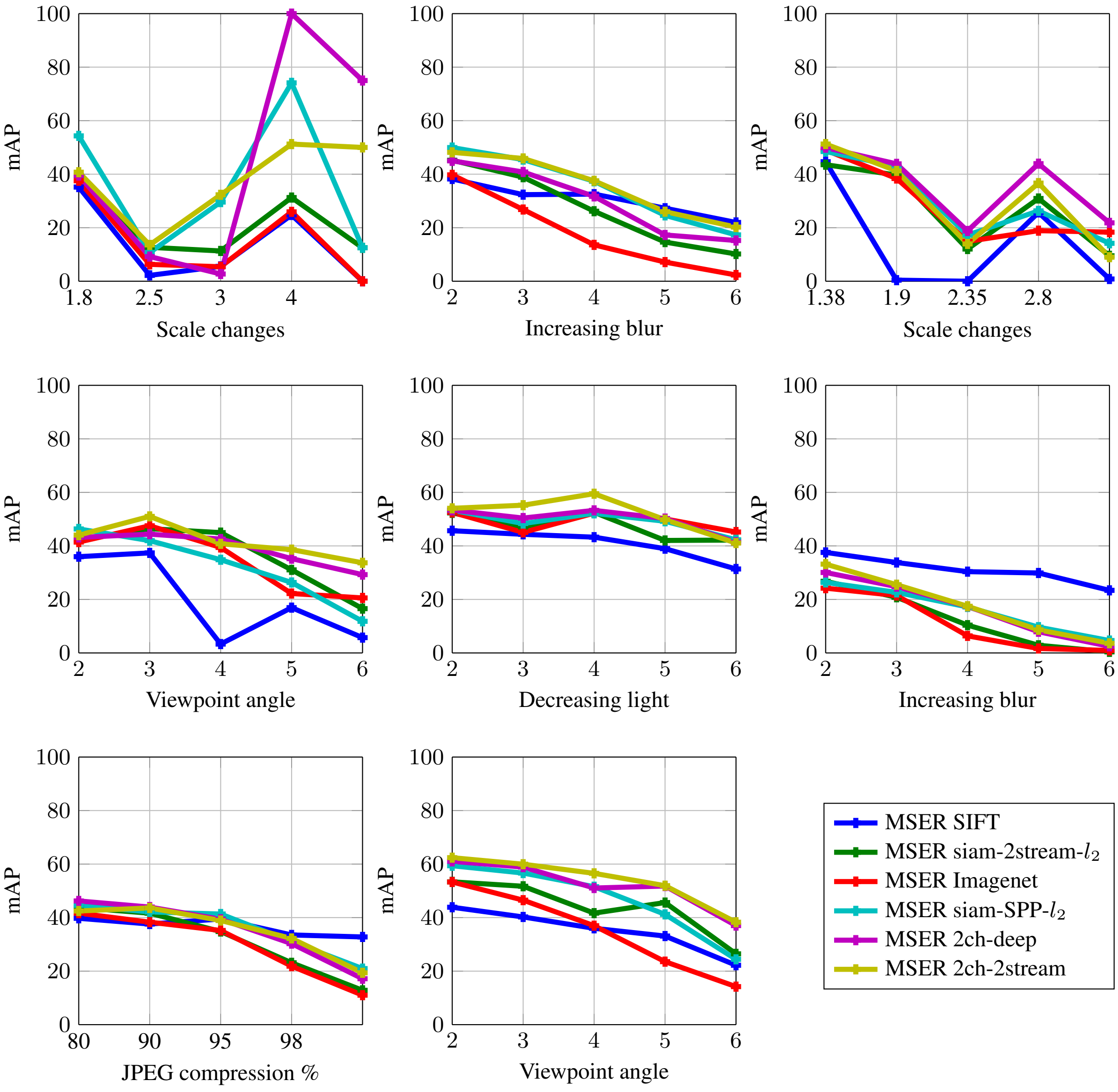


Figure 12. Evaluation plots of local descriptors on different datasets (*i.e.*, with different transformations). Horizontal axis represents the transformation magnitude in each case.

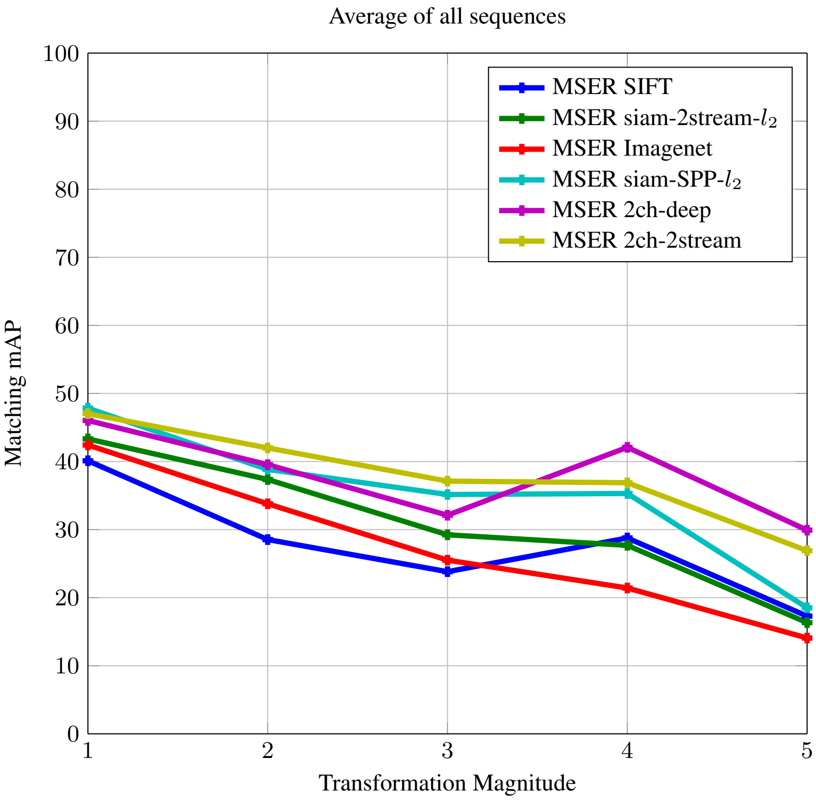


Figure 13. Overall evaluation of local descriptors showing the average performance over all datasets in Fig. 12.

3.1. SPP-based networks

We also experimented with evaluating the performance of SPP-based networks when using SPP layers of different spatial sizes. Minimal area size of detected with MSER ellipses set to 100. The results in fig. 14 concern the model siam-SPP-l_2 (recall that siam-SPP is obtained using siam descriptors, with spatial max-pooling module inserted after the second convolutional layer). The input patches were rescaled such that $\min(\text{width}, \text{height}) > a$ where a is a minimal image size accepted by the network and were equal to 34, 40, 46 and 64 for 1×1, 2×2, 3×3 and 4×4 spatial pooling output sizes respectively. Fig. 15 shows average mAP of all datasets. The results show that increasing pooling output size consistently improves results. It has to be noted that increasing pooling output leads to increased dimensionality of the descriptor, for example, 4x4 output size produces $192 \times 4 \times 4 = 3072$ dimensional feature. SPP performance can improve even further, as no multiple aspect ratio patches were used during training (these appear only at test time).

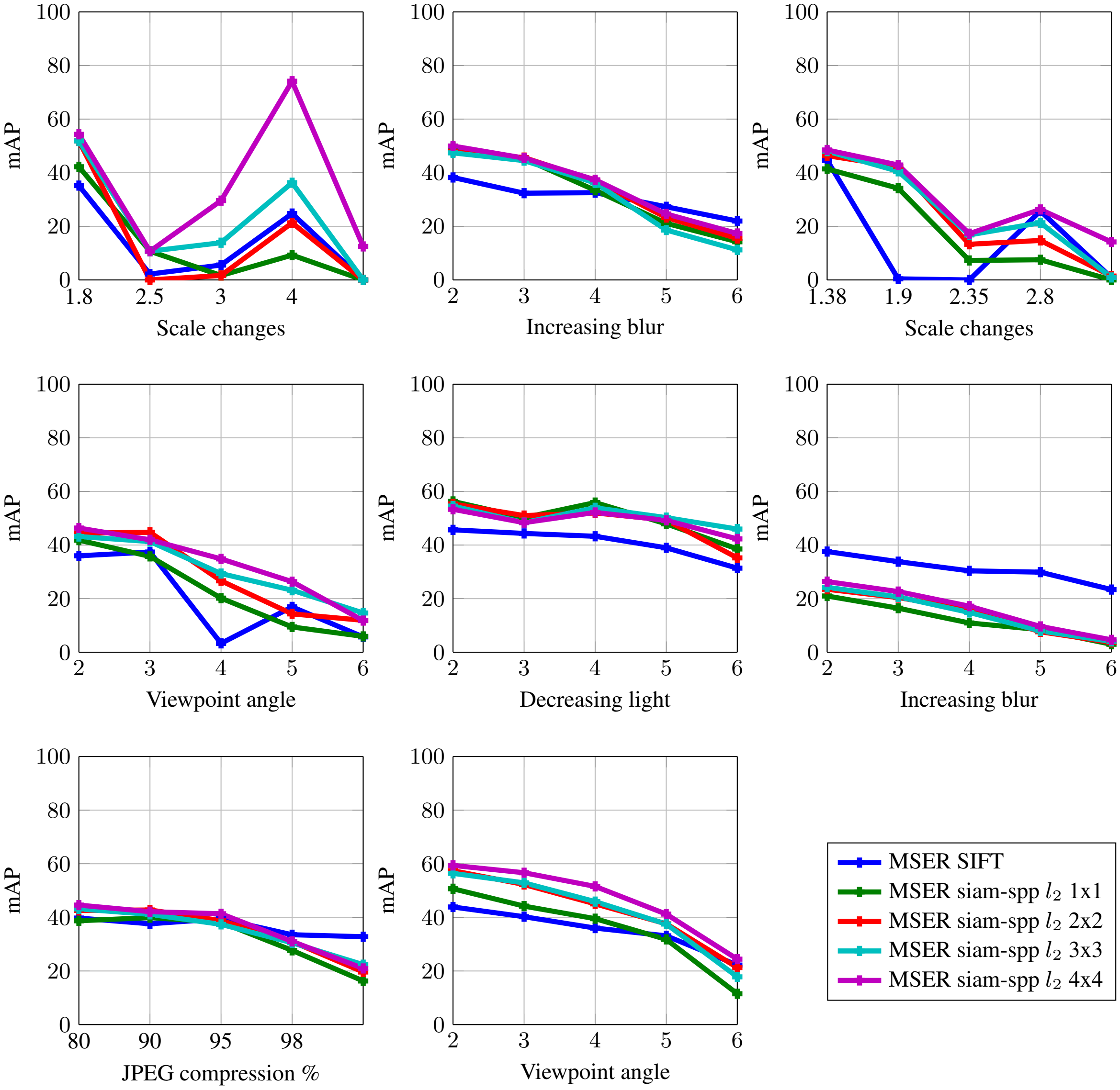


Figure 14. Evaluation plots of SPP-based network on different datasets when using SPP layers with different spatial sizes.

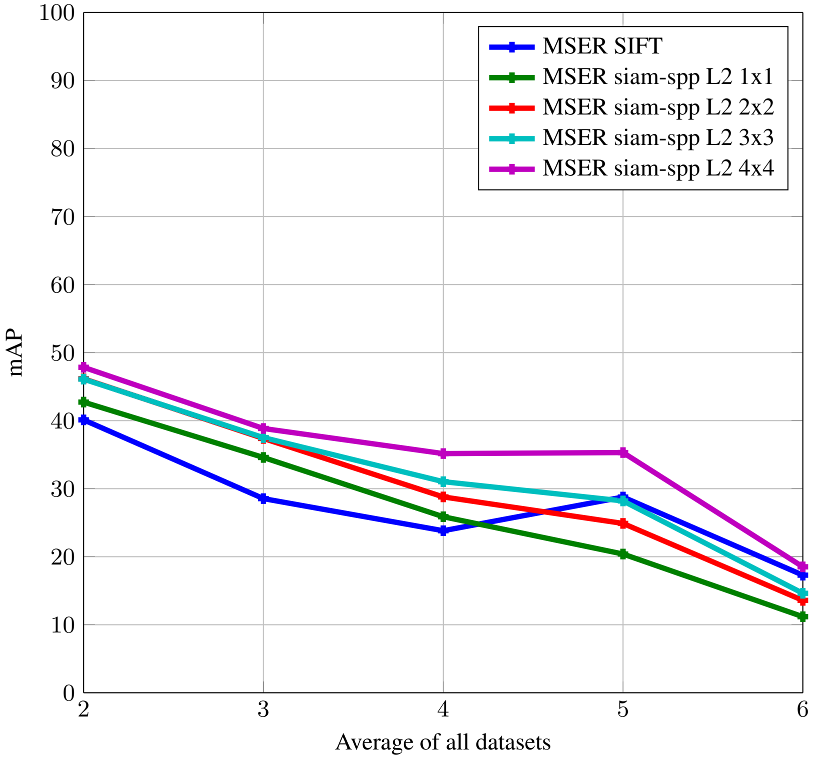


Figure 15. Overall performance when using SPP layers with different spatial sizes. We show average of all datasets of Fig. 14.

References

[1] G. H. M. Brown and S. Winder. Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 1

[2] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005. 1, 13

[3] K. Simonyan, A. Vedaldi, and A. Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 1

[4] C. Strecha, W. von Hansen, L. J. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *CVPR*. IEEE Computer Society, 2008. 3

[5] E. Tola, V.Lepetit, and P. Fua. A Fast Local Descriptor for Dense Matching. In *Proceedings of Computer Vision and Pattern Recognition*, Alaska, USA, 2008. 3