In this document, we provide additional explanations and results that were not included in the main text.

1. Method

1.1. Stability score

The process of ground truth generation is illustrated in Fig. 1. The restriction of deformations along the x and y axes, which is mentioned in the main text, is controlled by a parameter d. The parameter describes the length of an edge of a square embedded at the center of a larger square of size pd (see Sec. 3.2 of the main text). When deformations along x and y are sampled, they cannot distort the larger square in a way that its corners end up inside the square described by the parameter d. See the implementation in the GitHub repository for more details; file *source/projective/homography.py*, function *sample_homography*, parameter *scale_factor*.

1.2. Neural stability score

We explore the influence of parameters t_{Shi} and d on the performance of NeSS-ST. We train the network with different t_{Shi} and d and test obtained models on the validation subset of IMC-PT [9].

Based on the results of Fig. 2, Fig. 3 and Fig. 4, we pick $t_{Shi} = 0.005$ and $d = 2.0$.

2. Experiments

During experiments, non-maximum suppression sizes, score thresholds and parameters that are related to the structure of the model, *e.g.* the number of scales in Key.Net [2], are set for each model according to the values provided by the authors.

2.1. Evaluation on HPatches

We report repeatability [15] under different pixel thresholds following [6]. We perform hyper-parameter fine-tuning for the homography estimation task [6, 21]. We use images left out from the test set [7] as a validation set to get a general picture of the dependency between mAA and hyper-parameters (see Fig. 6). Because the validation set of HPatches consists only of 48 unique images and 40 image pairs, using the best parameters obtained on it during the evaluation on the test set gives inadequate results. To overcome the problem of a small validation set size we firstly take Lowe ratio parameters fine-tuned on IMC-PT [9] (see Table 2). We reason it is a good approximation since IMC-PT also has strong viewpoint changes like HPatches. Next, we notice that for some models increasing the inlier threshold up to a limit leads to overall improvements, hence, we select a high inlier threshold for all models. This decision is based on the fact that for other datasets inlier thresholds are not far off from each other (see Table 2 and Table 3). We estimate homographies using an OpenCV [3] routine with 10000 iterations and a 0.9999 confidence level.

Methods	Lowe ratio	Inlier threshold
Shi-Tomasi [8, 19] + DISK [20]	0.99	5.9
SIFT [13] + DISK [20]	0.98	6.0
SuperPoint [6] + DISK [20]	0.98	5.4
R2D2 [18] + DISK [20]	0.98	5.4
Key.Net [2] + DISK [20]	0.99	5.9
DISK [20]	0.98	5.4
REKD [11] + DISK [20]	0.98	5.2
NeSS-ST + DISK [20]	0.99	5.4

Table 1: Hyper-parameters used for homography estimation on HPatches [1].

Fig. 5 shows that our method has much lower repeatability compared to Shi-Tomasi [8, 19], however, keeps on par with it on the homography estimation task. Hyper-parameters reported in Table 1 provide significantly better results for all models among those that we tried on the test set. Although this kind of fine-tuning procedure is less principled than the one that we employ for IMC-PT [9] and ScanNet [5], we believe that it is the best option available given that HPatches doesn't have enough data to allow a proper hyper-parameter tuning.

2.2. Evaluation on downstream tasks

2.2.1 Evaluation on IMC-PT

We provide accuracy-threshold plots for the evaluation on a downstream task of relative pose estimation on IMC-PT [9] from which we calculate mAA [22, 9]. We use the validation subset of IMC-PT for hyper-parameter tuning. The set consists of 274 unique images that form 11.5k image pairs. In accordance with the protocol from [9], we first fine-tune the Lowe ratio, then - the inlier threshold. We plot mAA curves for different values of hyper-parameters for both rotation and translation. The best parameter is selected according to the largest sum of mAA for rotation and translation. We estimate the fundamental matrix by employing a robust estimator with DEGENSAC [4] using 200000 iterations and a 0.9999 confidence level. Additionally, we conduct experiments with different numbers of keypoints: we pick regimes of 128 and 512 keypoints to assess the ability of detectors to operate with a limited number of keypoints.

Fig. 7a shows that NeSS-ST consistently outperforms other self-supervised approaches over all thresholds. Fig. 8 illustrates the dependency between the hyper-parameters and mAA. We report hyper-parameters selected for each model. We found that slightly changing hyper-parameters

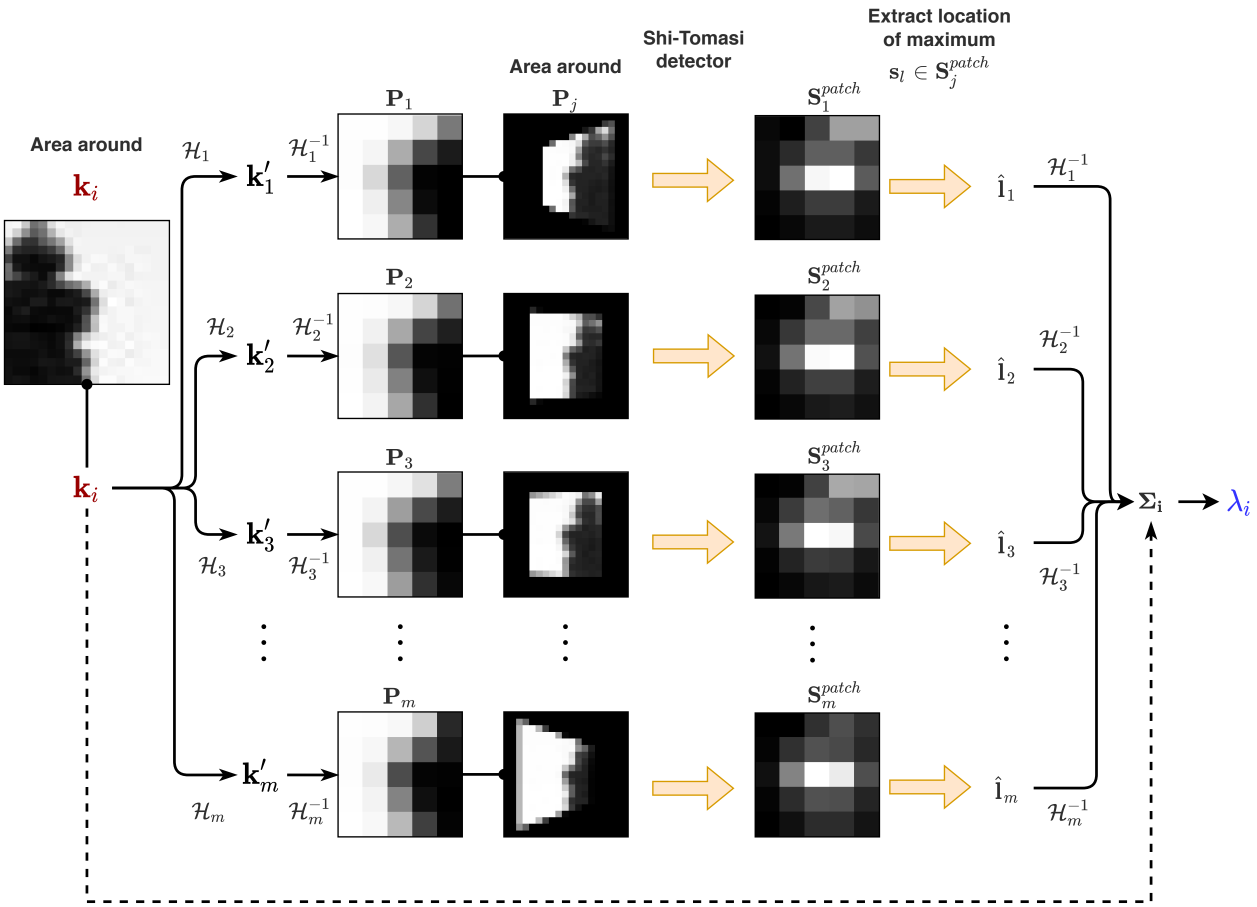


Figure 1: For each selected point $\mathbf{k}_i$ we calculate the ground truth stability score λ_i. Firstly, we generate a set of deformed patches $\{\mathbf{P}_j\}_{j=1}^m$ and run the Shi-Tomasi detector on patches to obtain a set of score patches $\{\mathbf{S}_j^{patch}\}_{j=1}^m$. For each $\mathbf{S}_j^{patch}$ we extract the location of its maximum score $\hat{\mathbf{l}}_j$ getting a set $\{\hat{\mathbf{l}}_j\}_{j=1}^m$. By transforming the elements of the set with $\{\mathcal{H}_j^{-1}\}_{j=1}^m$, we estimate $\mathbf{\Sigma}_i$ and calculate λ_i.

for some models improves their results on the test set, hence, the parameters used during the evaluation are slightly different from those selected on the validation set, see Table 2. Fig. 9a and Fig. 9b show that our method doesn't deal well with the decreased number of points. Firstly, we believe that it is related to our training setup where we select 1024 points per image. Secondly, in a few-point scenario points with higher repeatability, which our method doesn't look for, appear to present a better choice. We believe that this flaw in our method can be remedied by adding a term to the loss function that encourages correct predictions for different numbers of points.

2.2.2 Evaluation on MegaDepth

We provide accuracy-threshold plots for the evaluation on the downstream task of relative pose estimation on

Methods	Lowe ratio	Inlier threshold
Shi-Tomasi [8, 19] + DISK [20]	0.99	0.6
SIFT [13] + DISK [20]	0.98	0.6
SuperPoint [6] + DISK [20]	0.98	1.0
R2D2 [18] + DISK [20]	0.99	1.1
Key.Net [2] + DISK [20]	0.99	1.1
DISK [20]	0.98	0.7
REKD [11] + DISK [20]	0.98	1.1(1.3)
NeSS-ST + DISK [20]	0.99	0.6

Table 2: Hyper-parameters used for fundamental matrix estimation on IMC-PT [9]. Tuned hyper-parameters, if different, are provided in brackets.

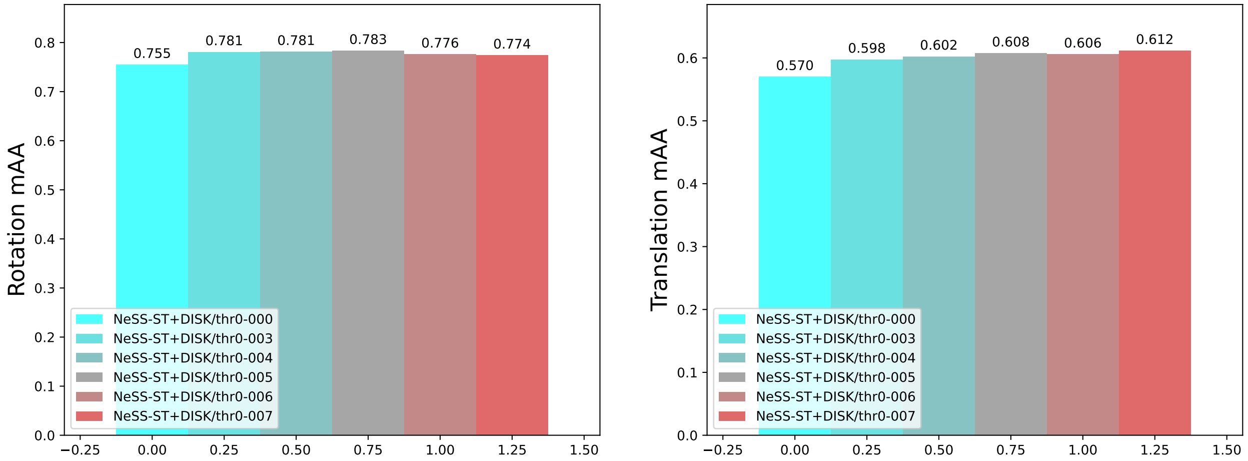


Figure 2: NeSS-ST models trained with different values of t_{Shi} (0.0, 0.003, 0.005, 0.006 and 0.007) evaluated on the validation set of IMC-PT [9] with 2048 keypoints and full resolution images. We report mAA [22, 9] up to a 10 degrees threshold for rotation and translation.

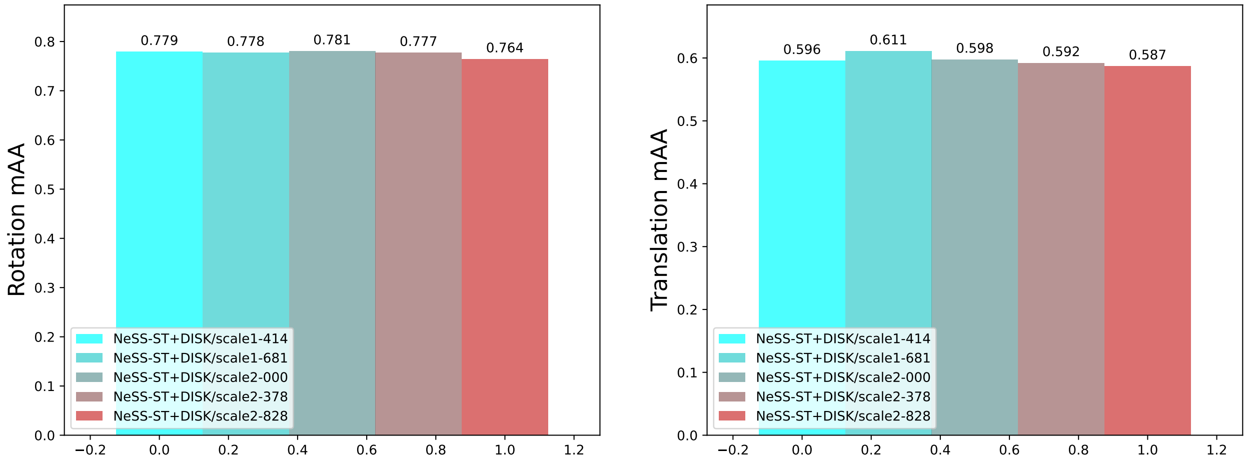


Figure 3: NeSS-ST models trained with different values of d (1.414, 1.681, 2.0, 2.378 and 2.828) evaluated on the validation set of IMC-PT [9] with 2048 keypoints and full resolution images. We report mAA [22, 9] up to a 10 degrees threshold for rotation and translation.

MegaDepth [12] from which we calculate mAA [22, 9].

Fig. 7b shows that NeSS-ST consistently outperforms all methods on translation estimation accuracy and is in the top three on rotation estimation accuracy over all thresholds.

2.2.3 Evaluation on ScanNet

We provide accuracy-threshold plots for the evaluation on the downstream task of relative pose estimation on ScanNet [5] from which we calculate mAA [22, 9]. We sample a validation set that consists of 16k unique images that form 8k pairs from validation sequences of ScanNet using a sampling procedure similar to [17, 21]. The hyper-parameter selection and fine-tuning are done in the same way as described in Sec. 2.2.1. We estimate the essential matrix using a robust estimator from OpenGV [10] with 5000 iterations and a 0.99999 confidence level.

Fig. 7c shows that NeSS-ST achieves second place over all thresholds losing only to R2D2 [18]. Fig. 10 presents the hyper-parameters and mAA curves. See Table 3 for the hyper-parameters used in the evaluation.

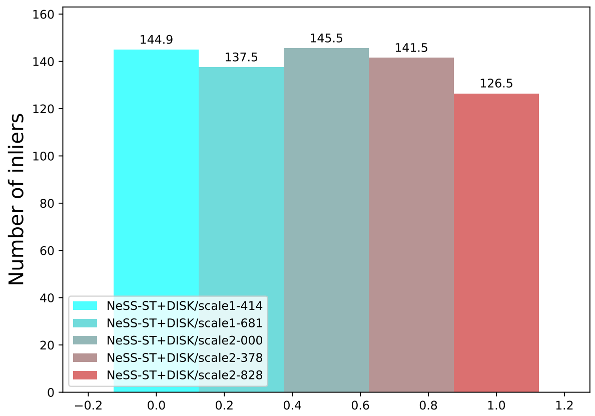


Figure 4: NeSS-ST models trained with different values of d (1.414, 1.681, 2.0, 2.378 and 2.828) evaluated on the validation set of IMC-PT [9] with 2048 keypoints and full resolution images. We report the number of inliers.

Methods	Lowe ratio	Inlier threshold
Shi-Tomasi [8, 19] + HardNet [16]	0.94	2.2
SIFT [13] + HardNet [16]	0.95	2.0
SuperPoint [6] + HardNet [16]	0.93	2.2
R2D2 [18] + HardNet [16]	0.92	2.4
Key.Net [2] + HardNet [16]	0.93	2.2(2.6)
DISK [20] + HardNet [16]	0.92	2.2
REKD [11] + HardNet [16]	0.89	2.2(2.4)
NeSS-ST + HardNet [16]	0.9	2.2

Table 3: Hyper-parameters used for essential matrix estimation on ScanNet [5]. Tuned hyper-parameters, if different, are provided in brackets.

2.3. Ablation study

2.3.1 Base detector ablation

We provide accuracy-threshold plots for the ablation on the downstream task of relative pose estimation on IMC-PT [9] from which we calculate mAA [22, 9]. Additionally, we report MMA [14, 7] and repeatability [15] for different pixel thresholds on HPatches.

Shi-Tomasi [8, 19] shows the best performance on the downstream task among all handcrafted detectors as illustrated in Fig. 11. In contrast with the downstream task evaluation, Fig. 12 shows that all handcrafted detectors have approximately the same performance in both MMA and repeatability. Using the detectors together with NeSS decreases their classical metrics scores, but, in return, it boots their performance on the downstream task. These results indicate that the conventional metrics cannot comprehensively assess the ability of a detector to perform well in applications.

2.3.2 Stability score design ablation

Like in Sec. 2.3.1, we provide mAA [22, 9] on IMC-PT [9] and classical metrics [14, 7] on HPatches [1]. Additionally, we conduct experiments to showcase the influence of thresholding on SS-ST and RS-ST detectors.

Fig. 13 shows that the neural network improves the performance of both SS-ST and RS-ST over all thresholds with models based on the stability score providing better downstream performance. Fig. 14b illustrates that RS-ST has higher repeatability than SS-ST. Still, the latter performs better on the downstream task. Fig. 15 illustrates that thresholding of low-saliency responses is essential for the good performance of both SS-ST and RS-ST.

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5173–5182, 2017.

[2] Axel Barroso-Laguna, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5836–5844, 2019.

[3] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[4] Ondrej Chum, Tomas Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 772–779. IEEE, 2005.

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017.

[6] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 224–236, 2018.

[7] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv preprint arXiv:1905.03561*, 2019.

[8] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the 4th Alvey Vision Conference*, pages 147–151, 1988.

[9] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image

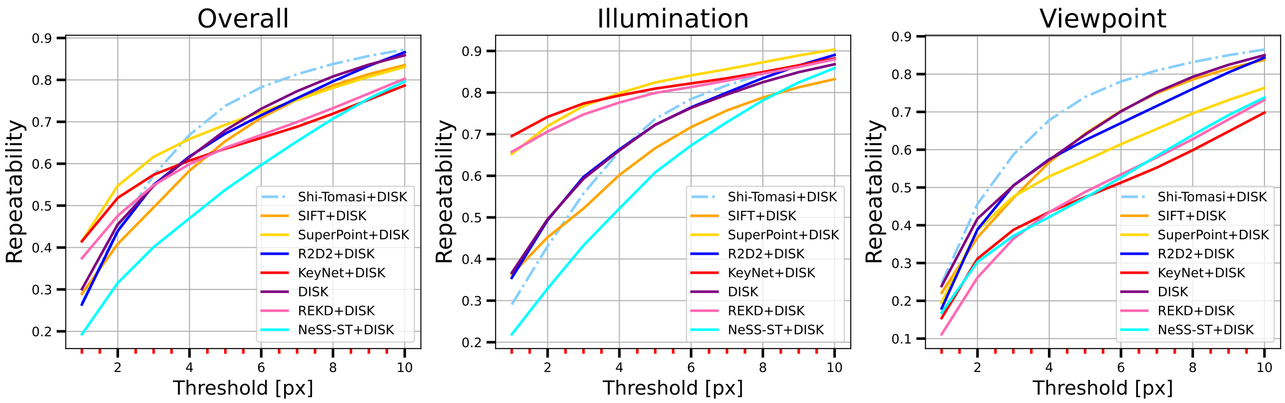
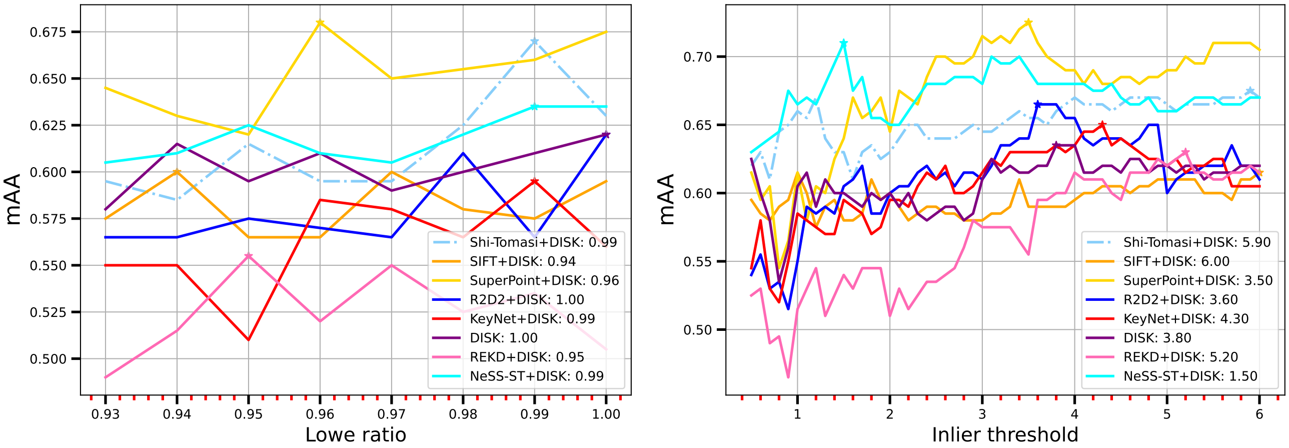


Figure 5: Evaluation on HPatches [1] with 2048 keypoints and full resolution images. We report repeatability [15].



(a) Lowe ratio tuning.

(b) Inlier threshold tuning.

Figure 6: Hyper-parameter tuning on the validation set of HPatches [1] with with 2048 keypoints and full resolution images. We report homography estimation mAA [6, 21, 22, 9] up to a 5-pixel threshold for different hyper-parameter values.

matching across wide baselines: From paper to practice. *Intl. J. of Computer Vision*, 129(2):517–547, 2021.

[10] Laurent Kneip and Paul Furgale. Opengv: A unified and generalized approach to real-time calibrated geometric vision. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2014.

[11] Jongmin Lee, Byungjin Kim, and Minsu Cho. Self-supervised equivariant learning for oriented keypoint detection. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4847–4857, 2022.

[12] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018.

[13] David G Lowe. Distinctive image features from scale-invariant keypoints. *Intl. J. of Computer Vision*, 60(2):91–110, 2004.

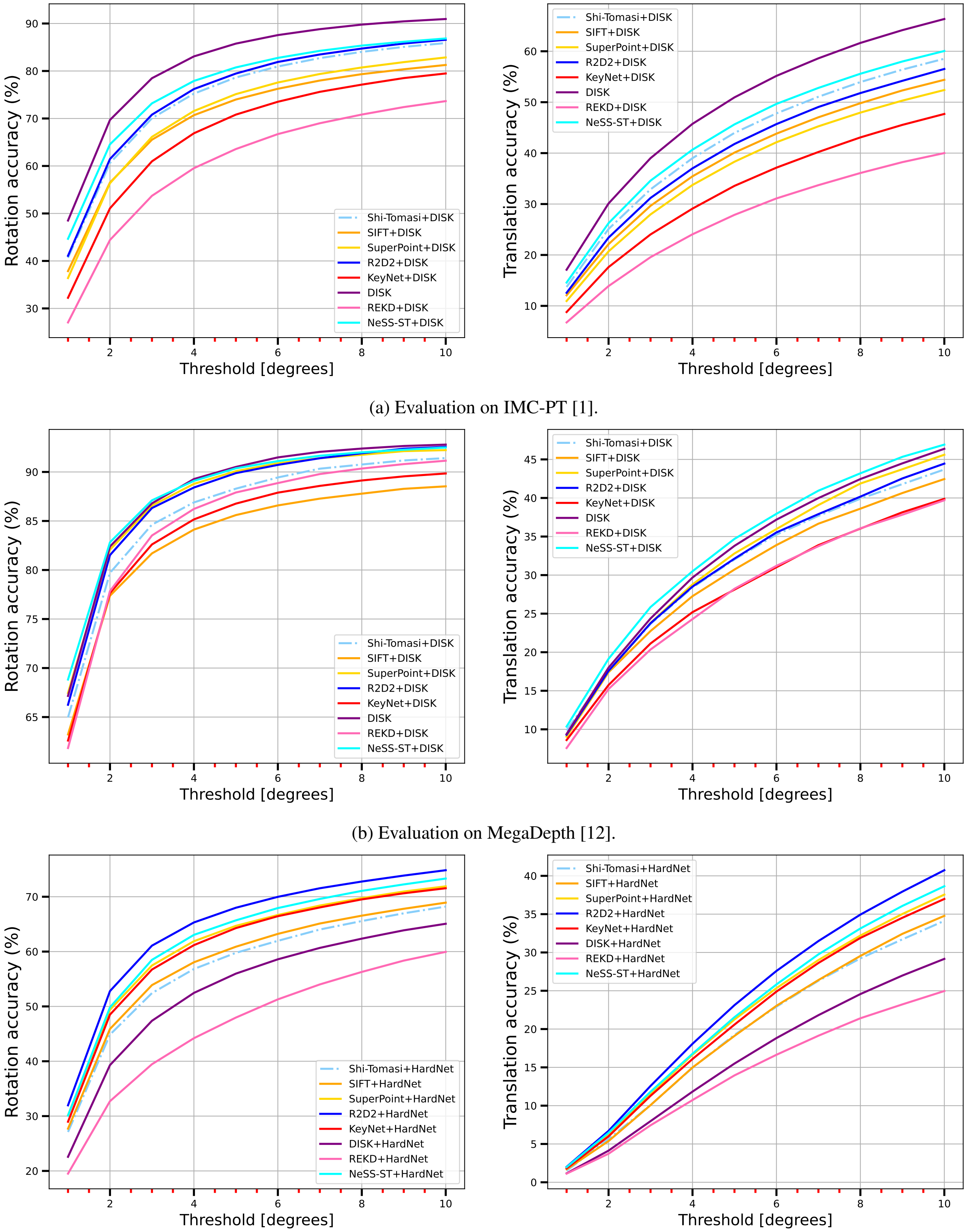
[14] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(10):1615–1630, 2005.

[15] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and L Van Gool. A comparison of affine region detectors. *Intl. J. of Computer Vision*, 65:43–72, 2005.

[16] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

[17] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning local features from images. *Advances in Neural Information Processing Systems (NIPS)*, 31, 2018.

[18] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. In H. Wallach, H. Larochelle, A. Beygelz-

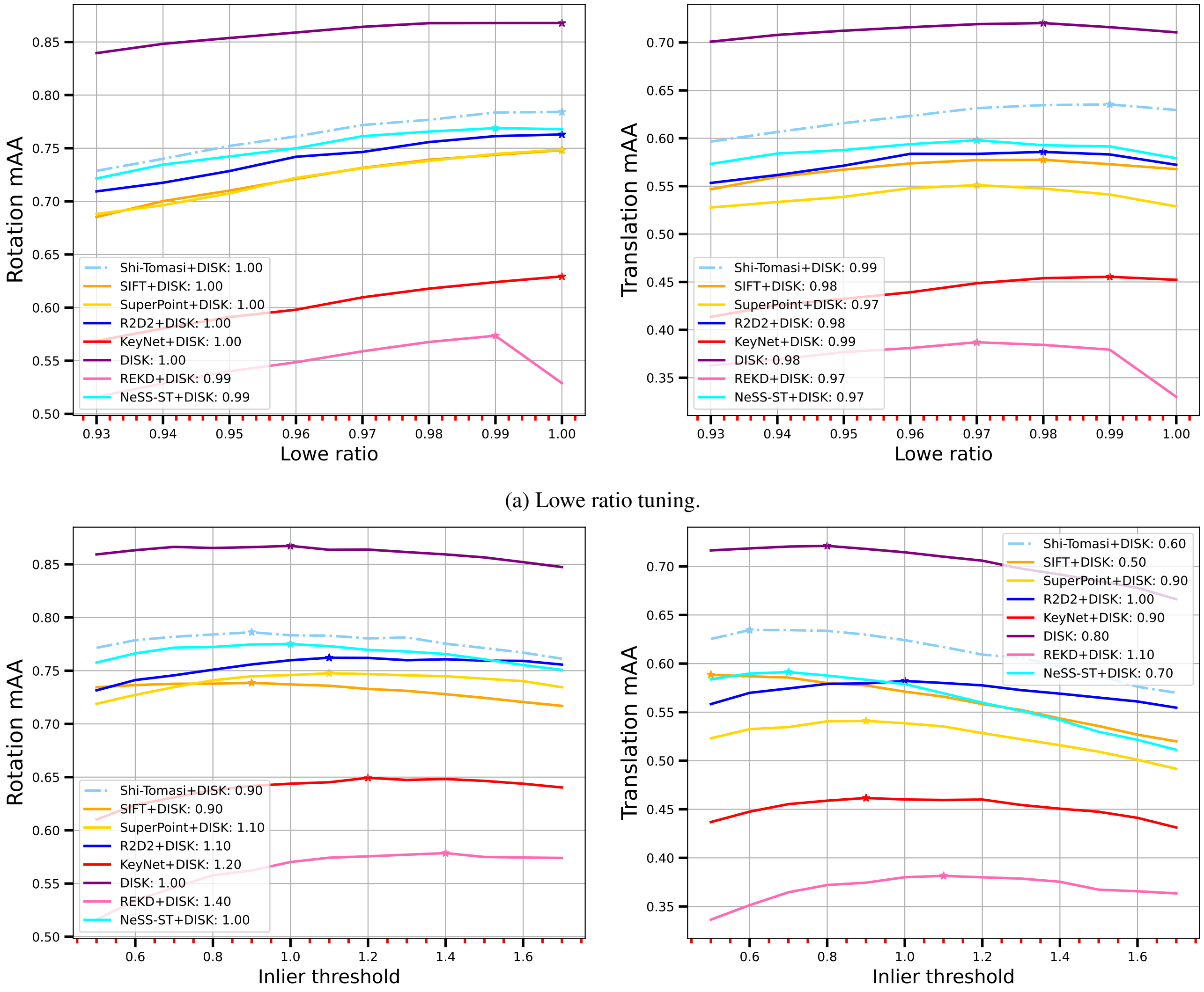


(a) Evaluation on IMC-PT [1].

(b) Evaluation on MegaDepth [12].

(c) Evaluation on ScanNet [5].

Figure 7: Evaluation on the downstream task of pose estimation with 2048 keypoints and full resolution images. We report pose estimation accuracy [22, 21, 9] in %.



(a) Lowe ratio tuning.

(b) Inlier threshold tuning.

Figure 8: Hyper-parameter tuning on the validation set of IMC-PT [9] with with 2048 keypoints and full resolution images. We report mAA [22, 9] up to a 10 degrees threshold for rotation and translation for different hyper-parameter values.

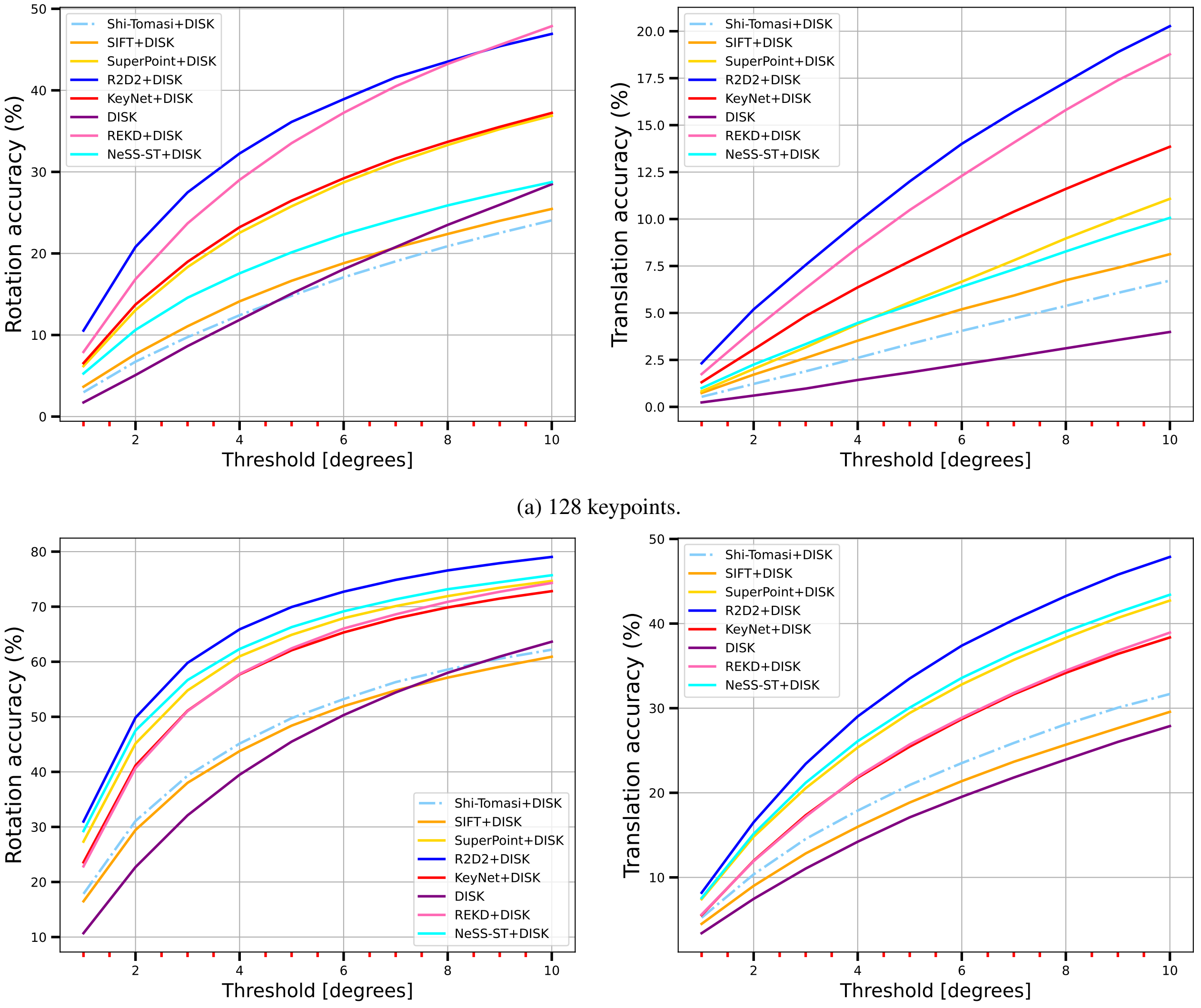
imer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 32. Curran Associates, Inc., 2019.

[19] Jianbo Shi and Tomasi. Good features to track. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 593–600, 1994.

[20] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems (NIPS)*, 33:14254–14265, 2020.

[21] Qianqian Wang, Xiaowei Zhou, Bharath Hariharan, and Noah Snavely. Learning feature descriptors using camera pose supervision. In *Eur. Conf. on Computer Vision (ECCV)*, pages 757–774. Springer International Publishing, 2020.

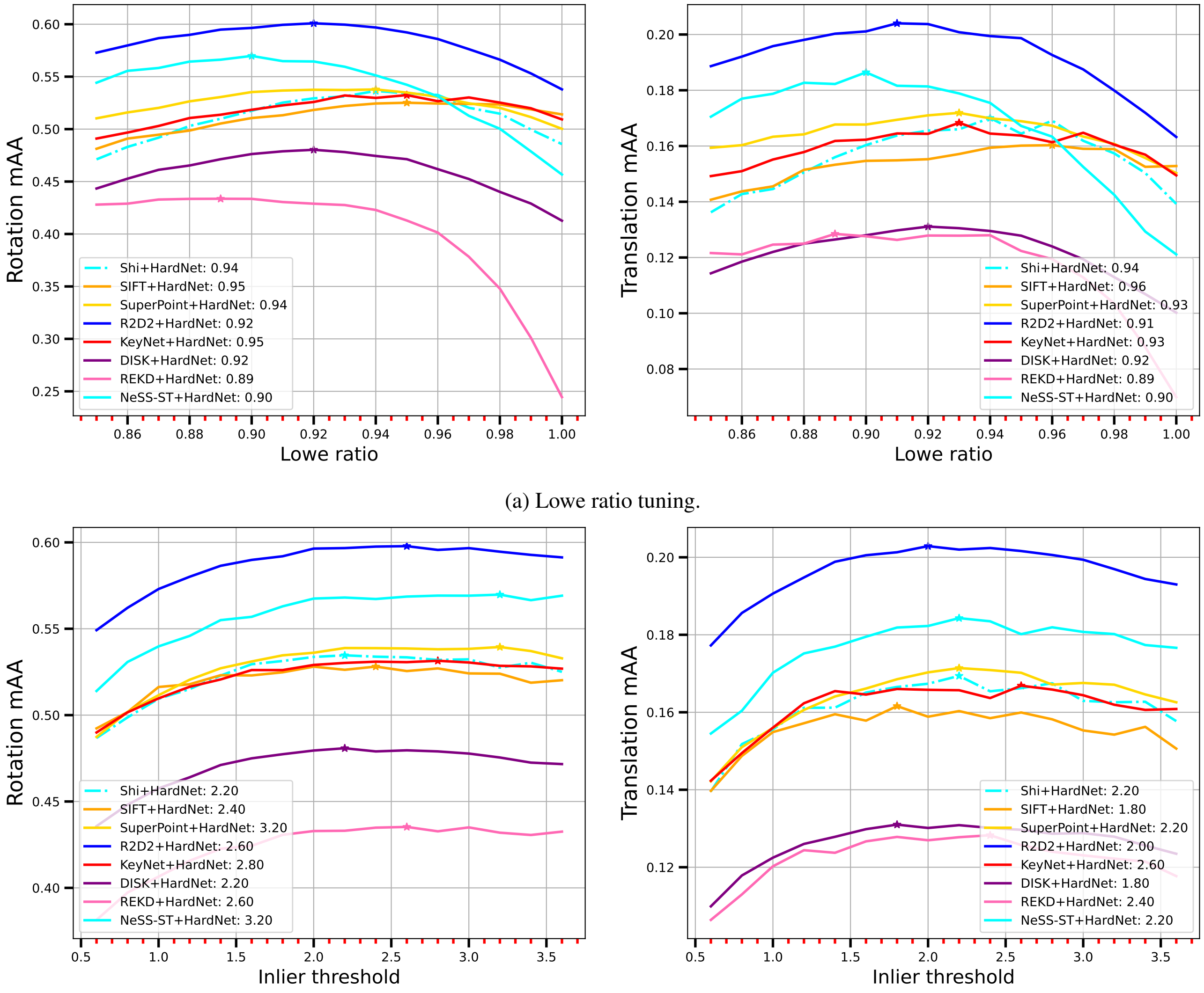
[22] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2666–2674, 2018.



(a) 128 keypoints.

(b) 512 keypoints.

Figure 9: Ablation on IMC-PT [1] with different numbers of keypoints and full resolution images. We report pose estimation accuracy [22, 21, 9] in %.



(a) Lowe ratio tuning.

(b) Inlier threshold tuning.

Figure 10: Hyper-parameter tuning on the validation set of ScanNet [5] with with 2048 keypoints and full resolution images. We report mAA [22, 9] up to a 10 degrees threshold for rotation and translation for different hyper-parameter values.

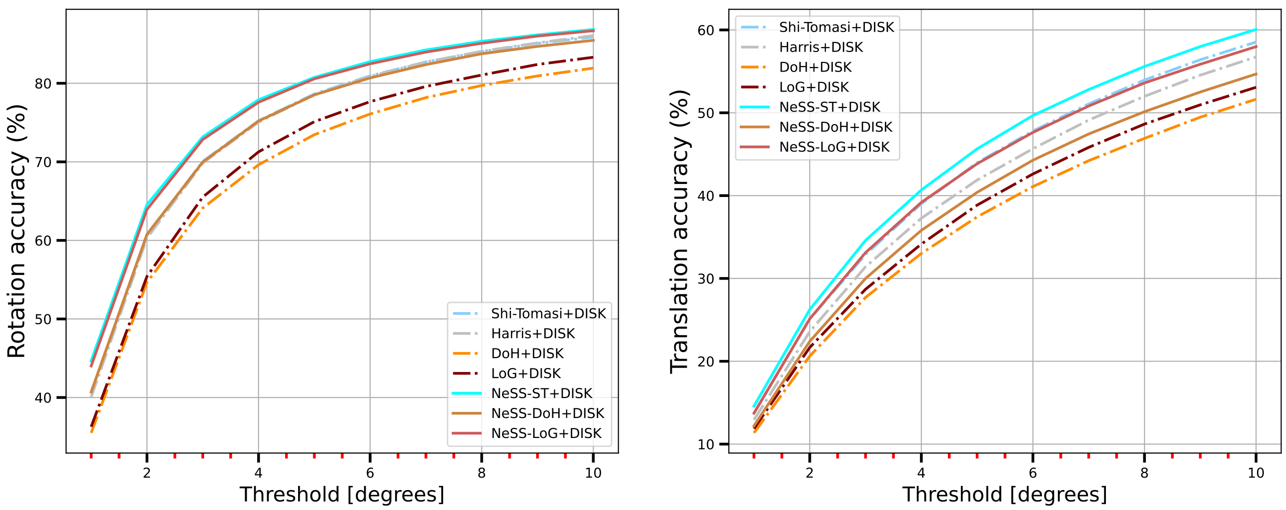
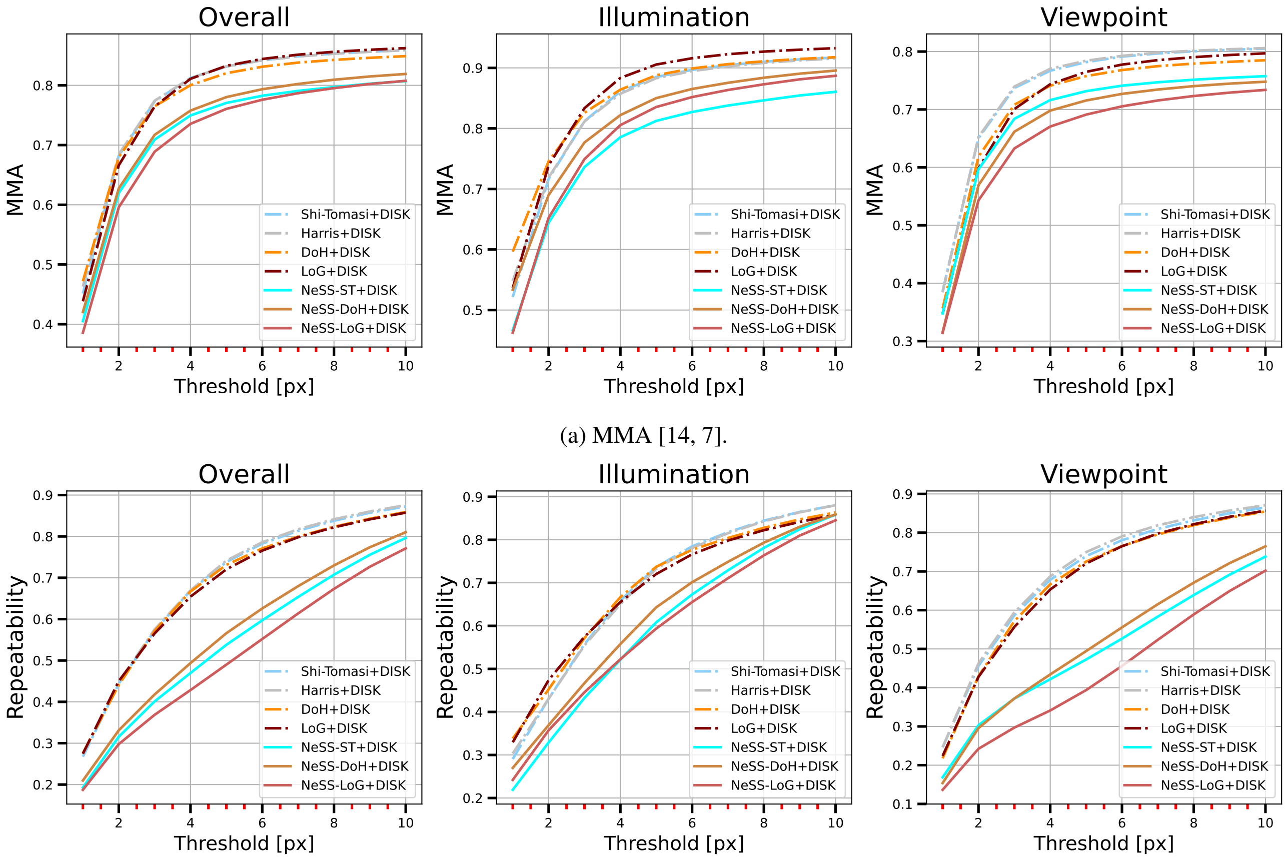


Figure 11: Base detector ablation on IMC-PT [9] with 2048 keypoints and full resolution images. We report pose estimation accuracy [22, 21, 9] in %.



(a) MMA [14, 7].

(b) Repeatability [15].

Figure 12: Base detector ablation on HPatches [1] with 2048 keypoints and full resolution images. We report classical metrics.

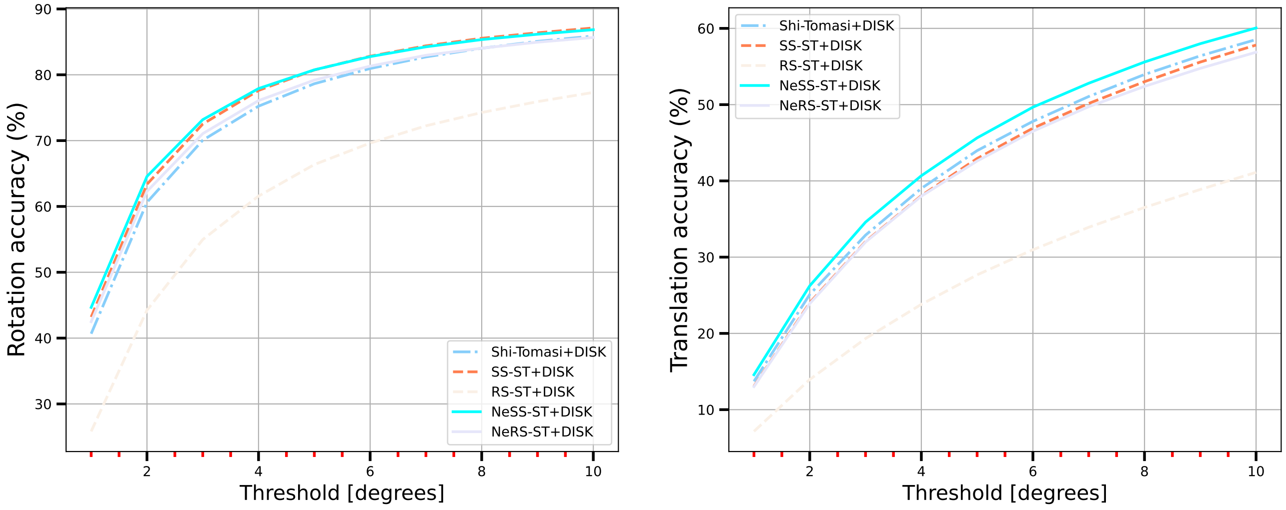
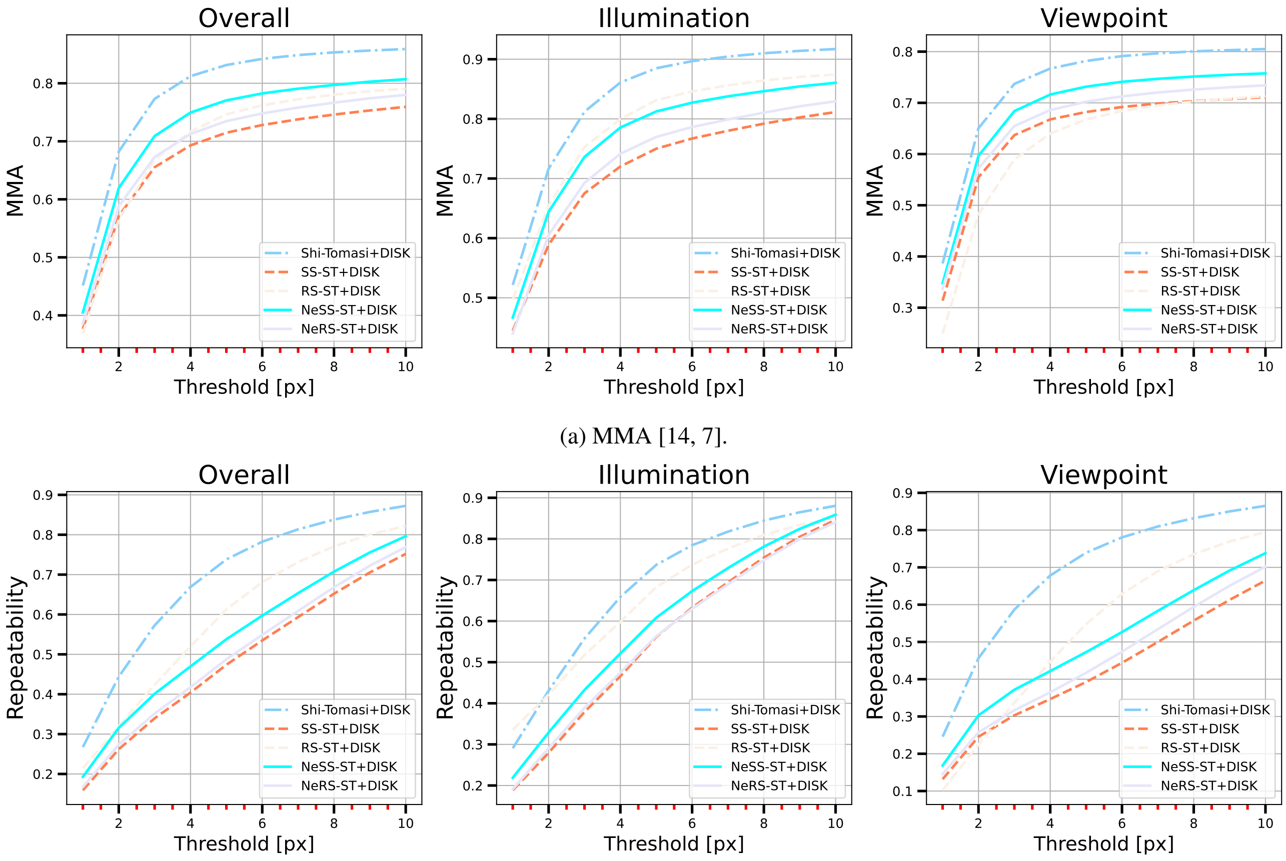


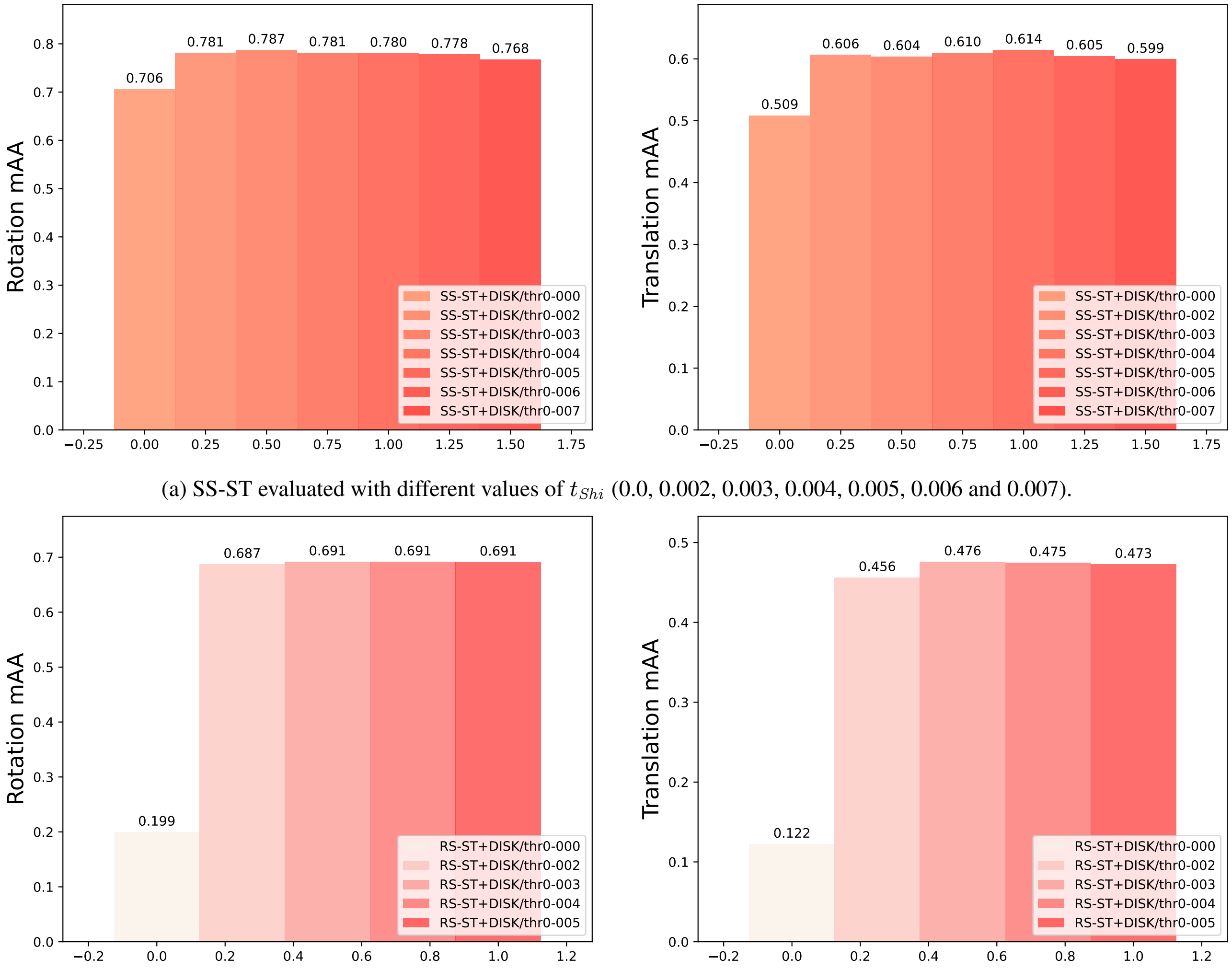
Figure 13: SS design ablation on IMC-PT [9] with 2048 keypoints and full resolution images. We report pose estimation accuracy [22, 21, 9] in %.



(a) MMA [14, 7].

(b) Repeatability [15].

Figure 14: SS design ablation on HPatches [1] with 2048 keypoints and full resolution images. We report classical metrics.



(a) SS-ST evaluated with different values of t_{Shi} (0.0, 0.002, 0.003, 0.004, 0.005, 0.006 and 0.007).

(b) RS-ST evaluated with different values of t_{Shi} (0.0, 0.002, 0.003, 0.004 and 0.005).

Figure 15: Ablation of the influence of filtering on performance on the validation set of IMC-PT [9] with 2048 keypoints and full resolution images. We report mAA [22, 9] up to a 10 degrees threshold for rotation and translation.