Supplementary Materials

1. Maths

In this section, we provide a short mathematical description of our method (probabilistic modeling and losses).

1.1. Definitions

- I, I' is the pair of images to match.
- q_i, $q'_j \in \mathbb{R}$ are the ith and jth keypoint logits from I and I' respectively (M in total).
- d_i, $d'_j \in \mathbb{R}^{128}$ are the ith and jth descriptors from I and I' respectively (M in total).
- $s_{ij} \in [-1, +1]$ is the similarity score between keypoint i from I and keypoint j from I'. s is therefore a $M \times M$ matrix.
- c_i, c'_j are the ith and jth correspondence indices from I and I' respectively (N in total, with $N \leq M$). Such that c_i and c'_i represent the correspondence between descriptor d_{c_i} and $d'_{c'_i}$ with similarity $s_{c_i c'_i}$.

1.2. Matching Probabilities

The matching probabilities are modeled by a double-softmax, enforcing the cycle-consistency property (cf. Fig. 1).

$$P_{i \leftrightarrow j} = P_{i \rightarrow j} P_{i \leftarrow j}$$

- $P_{i \rightarrow j} = \frac{e^{\frac{s_{ij}}{\tau}}}{\sum_k e^{\frac{s_{ik}}{\tau}}}$ is the directional probability of matching d_i to d'_j
- $P_{i \leftarrow j} = \frac{e^{\frac{s_{ij}}{\tau}}}{\sum_k e^{\frac{s_{kj}}{\tau}}}$ is the directional probability of matching d'_j to d_i

where τ is the temperature, and s the pairwise similarity matrix; obtained using a standard cosine similarity function.

$$s_{ij} = \text{cosim}(d_i, d'_j) = \frac{\langle d_i, d'_j \rangle}{\sqrt{\langle d_i, d_i \rangle \langle d'_j, d'_j \rangle}}$$

where $\langle ., . \rangle$ is the dot product.

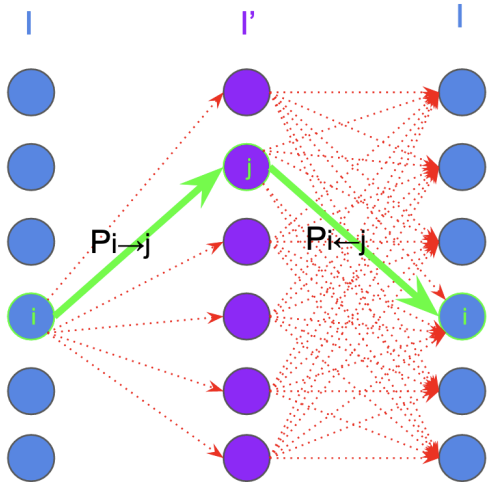

Figure 1: Visualization of the cycle-consistent probabilistic path. The probability $P_{i \leftrightarrow j}$ is the probability of following the green path, over the set of all possible red paths; from image I to I' and back.

1.3. Keypoint Probabilities

All keypoints probabilities are obtained using a simple sigmoid.

$$\sigma(q_i) = \frac{1}{1 + e^{-q_i}}$$

1.4. Matching Loss

The matching loss of a single image pair is the negative log likelihood loss, summed over the entire set of correspondences.

$$\begin{aligned} \mathcal{L}_{desc} &= \text{NLL}(s, c, c') \\ &= -\frac{1}{N} \sum_{i=0}^{N-1} \log P_{c_i \leftrightarrow c'_i} \\ &= -\frac{1}{N} \sum_{i=0}^{N-1} \left[\log P_{c_i \rightarrow c'_i} + \log P_{c'_i \leftarrow c_i} \right] \end{aligned}$$

1.5. Keypoint Loss

Once the matching success of descriptors as been measured (and stored in variable $y \in \{0,1\}^N$, cf. Sec. 1.6), we can learn the keypoint probabilities using a standard binary cross-entropy loss; using the keypoint logits extracted from both I and I'.

$$\mathcal{L}_{key} = \text{BCE}(q, y, c) + \text{BCE}(q', y, c')$$

where

$$\text{BCE}(q, y, c) = -\frac{1}{N}\sum_{i=0}^{N-1}\Big[y_i \log \sigma(+q_{c_i}) + (1-y_i)\log\sigma(-q_{c_i})\Big]$$

1.6. Matching Success

The matching success is the process of measuring whether or not matching currently learned descriptors (using a simple mutual nearest neighbor) would produce correct matches. It can be expressed mathematically as verifying whether or not the similarity of a ground truth correspondence is the row and column maximum in s.

$$y_i = \mathbf{1}\big[s_{c_i c'_i} \geq \max_k\{s_{c_i k}\}\big]\mathbf{1}\big[s_{c_i c'_i} \geq \max_k\{s_{k c'_i}\}\big]$$

where $\mathbf{1}\big[.\big]$ is the indicator function.

2. Additional Experiments

In this section, we present additional experiments as evidence of the robustness of SiLK under varying, but realistic, conditions. We also hope this comprehensive set of data points can be used by the community to tune their own version of SiLK to a specific use case or task. For example, if a task is sensitive to false positive matching, one might consider using the ratio-test filtering, as indicated in Tab. 3.

SiLK's robustness is tested against three important types of variations, as we aim to answer the following.

1) Are SiLK's results robust across different image resolutions ? (cf. Tab. 1 and Fig. 2)

2) Can we improve SiLK by simply selecting more keypoints ? Or do we reach saturation for a certain value of k ? (cf. Tab. 2 and Fig. 3)

3) Do existing false-positive removal techniques work on SiLK? And how is performance affected by it ?(cf. Tab. 3).

Additionally, we empirically demonstrate the importance of **not** using zero-padding when learning keypoints (cf. Tab. 4). This is an often under-emphasized point we shed light on here.

2.1. Downsizing images is better than increasing ϵ

All of the HPatches metrics reported in this paper use an ϵ-distance error threshold to determine whether or not a pixel position is considered close-enough to its ground truth. A low ϵ means the metrics are reported in a highly accurate regime, while a high value of ϵ allows for some local mistakes to occur. However, ϵ is an absolute pixel distance, which means that metrics might vary in non-trivial ways as the input resolution changes during inference.

Existing methods tend to report metrics using high values of ϵ. For example, D2-Net [1], R2D2 [3] and DISK [5] all report MMA with ϵ-thresholds up to 10. We argue this is unnecessary. Tasks that do not require accurate keypoints (i.e. high values of ϵ) might want to reconsider running their keypoint model on lower resolution images (to reduce computational cost). As an input image is downscaled by a factor γ, ϵ should also be downscaled by the same factor in order to keep its relative size constant. So in theory, a metric reported on resolution α with error-threshold ϵ should be roughly equivalent to the same metric reported on resolution $\frac{\alpha}{\gamma}$ with error threshold $\frac{\epsilon}{\gamma}$.

In Tab. 1, we show this initial intuition is not correct. When looking at resolutions of 720 and 240 (i.e. downscaling of γ=3), existing methods all underperform their expected theoretical values. SiLK is the only model that consistently gain from running at lower resolutions. This is likely caused by SiLK's ability to obtain high number of keypoints on low resolution images while other methods are limited by their sparsity constraints (i.e. cell-based keypoint detection and NMS).

Additionally, we show in Fig. 2 that SiLK's performance against sparse keypoint methods is robust across a wide range of resolutions; with the exception of SIFT on large image resolution, using Homography Estimation Accuracy metric. One can also observe large performance gain (except on MMA) versus LoFTR [4](dense keypoint method) in the low resolution regime.

2.2. Increasing top-k improves SiLK, but saturates early

Increasing top-k has shown multiple times (cf. Tab. 1) to improve overall results. In this experiment, we simply vary the parameter k in order to monitor SiLK's performance.

As can be observed in Fig. 3 and Tab. 2, results start to saturate after $k = 10,000$; on Homography Estimation and MMA. Repeatability continuously increases as we increase k, but that is simply a consequences of getting more keypoints (i.e. the keypoint overlaps become more likely).

2.3. Improving MMA using ratio-test or double-softmax filtering

All previously reported results have been computed using MNN matching on unprocessed cosine distances. There

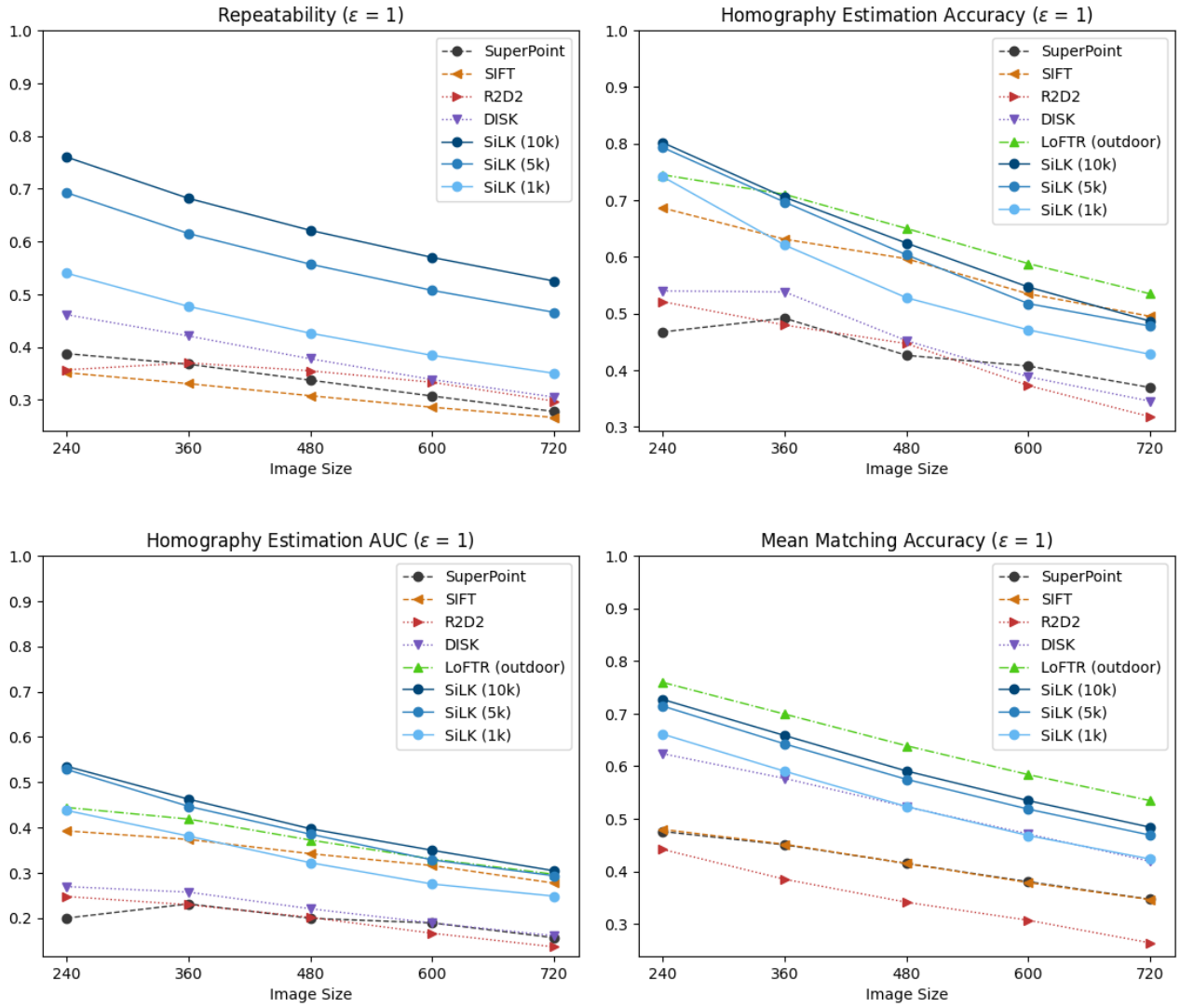

Figure 2: **SiLK's performance is robust across different input scales.** All sparse methods are outperformed by SiLK on all metrics and all scales (except SIFT on Homography Estimation Accuracy and large image resolution). Notice how R2D2, DISK and SuperPoint rankings differ across resolutions.

are, however, known distance post-processing techniques used to reduce false positive matching. The ratio-test [2] is one of those. The distance of the best match is divided by the distance of the second best match. A low value indicates a large difference between the two best distances, which indicates a measure of relative distinctiveness. Therefore, filtering out matches with high ratio values do tend to reduce matching errors caused by repeated similar keypoints (e.g. window corners of a building).

More recently[4], a similar idea has emerged from the probabilistic formulation of the matching problem: Filtering out low-probability matches seems like a natural way to reduce false positive matches.

In Fig. 4 and Tab. 3, we show that using either ratio-test or double-softmax filtering can help SiLK trade Homography Estimation for MMA.

2.4. Use NO padding to learn keypoints.

Zero padding is commonly used in various models. It is the process of adding a 0-filled border to an image or dense feature map. A common example is when using 3x3 convolutions, adding a padding of 1 will ensure the spatial shape of the input is preserved, otherwise it would be reduced.

The use or lack of padding is rarely mentioned by exist-

HPatches											
	Size	Repeatability		Hom. Est. Acc.		Hom. Est. AUC		MMA		# of keypoints	
		$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	pre-match	post-match
SuperPoint	240	0.39×	0.64	0.47×	0.83	0.20×	0.54	0.48×	0.76	305	187
	360	0.37	0.63	0.49	0.83	0.23	0.55	0.45	0.74	569	341
	480	0.34	0.61	0.43	0.80	0.20	0.51	0.41	0.72	847	499
	600	0.31	0.59	0.41	0.79	0.19	0.49	0.38	0.69	1141	655
	720	0.28	0.56✓	0.37	0.74✓	0.16	0.45✓	0.35	0.66✓	1460	813
SIFT	240	0.35×	0.54	0.69×	0.89	0.39×	0.68	0.48×	0.57	677	298
	360	0.33	0.53	0.63	0.85	0.37	0.65	0.45	0.56	1331	570
	480	0.31	0.52	0.60	0.84	0.34	0.61	0.41	0.55	2189	910
	600	0.29	0.51	0.53	0.80	0.32	0.58	0.38	0.53	3212	1297
	720	0.27	0.50✓	0.49	0.75✓	0.28	0.53✓	0.35	0.50✓	4273	1681
R2D2	240	0.36×	0.70	0.52×	0.81	0.25×	0.56	0.44×	0.79	1037	351
	360	0.37	0.73	0.48	0.82	0.23	0.54	0.38	0.77	3193	1042
	480	0.36	0.72	0.45	0.79	0.20	0.50	0.34	0.75	6088	1967
	600	0.33	0.71	0.37	0.76	0.17	0.46	0.31	0.72	9517	2994
	720	0.30	0.68✓	0.32	0.69✓	0.14	0.42✓	0.26	0.67✓	12036	3698
DISK	240	0.46×	0.72	0.54×	0.86	0.27×	0.60	0.62×	0.87	841	484
	360	0.42	0.71	0.54	0.85	0.26	0.58	0.58	0.86	1847	1030
	480	0.38	0.69	0.45	0.80	0.22	0.52	0.52	0.84	3349	1794
	600	0.34	0.67	0.39	0.75	0.19	0.47	0.47	0.81	5152	2647
	720	0.31	0.65✓	0.34	0.71✓	0.16	0.43✓	0.42	0.77✓	7417	3732
LoFTR (outdoor)	240	-	-	0.74×	0.90	0.44×	0.72	0.76×	0.93	1280	662
	360	-	-	0.71	0.90	0.42	0.70	0.70	0.92	2878	1533
	480	-	-	0.65	0.87	0.37	0.65	0.64	0.91	5109	2719
	600	-	-	0.59	0.83	0.33	0.61	0.58	0.89	7987	4166
	720	-	-	0.53	0.79✓	0.30	0.57✓	0.53	0.87✓	11490	5804
SiLK (top-10k)	240	0.76✓	0.90	0.80✓	0.93	0.54✓	0.78	0.73✓	0.79	10000	4816
	360	0.68	0.85	0.71	0.91	0.46	0.72	0.66	0.75	10000	4515
	480	0.62	0.81	0.62	0.87	0.40	0.66	0.59	0.71	10000	4283
	600	0.57	0.77	0.55	0.81	0.35	0.59	0.53	0.67	10000	4092
	720	0.53	0.73×	0.49	0.76×	0.30	0.53×	0.48	0.63×	10000	3945
SiLK (top-5k)	240	0.69✓	0.86	0.79✓	0.93	0.53✓	0.77	0.71✓	0.77	5000	2331
	360	0.62	0.80	0.70	0.90	0.45	0.70	0.64	0.73	5000	2181
	480	0.56	0.76	0.60	0.85	0.39	0.64	0.57	0.69	5000	2074
	600	0.51	0.71	0.52	0.80	0.33	0.57	0.52	0.65	5000	1983
	720	0.47	0.67×	0.48	0.74×	0.29	0.52×	0.47	0.61×	5000	1919
SiLK (top-1k)	240	0.54✓	0.73	0.74✓	0.91	0.44×	0.72	0.66✓	0.71	1000	429
	360	0.48	0.66	0.62	0.86	0.38	0.65	0.59	0.67	1000	408
	480	0.43	0.61	0.53	0.81	0.32	0.58	0.52	0.63	1000	389
	600	0.38	0.57	0.47	0.75	0.28	0.52	0.47	0.59	1000	376
	720	0.35	0.53×	0.43	0.69×	0.25	0.48✓	0.42	0.55×	1000	366

Table 1: **SiLK is the only model that benefits from running at lower resolutions.** On each metric and method, we compare the lowest (resolution=240, ϵ=1) pair, to the highest (resolution=720, ϵ=3) pair. Since the resolution / ϵ ratio is the same for both pairs, the measured level of accuracy is equivalent.

ing keypoint methods. However, we find that most implementations do in fact avoid the use of zero padding. Other implementations might use it, but then compensate by removing an arbitrarily-sized border from the dense outputs.

The reason for not using padding in SiLK's case is because it creates easily detectable corners and edges on the image borders, therefore causing overfitting during training.

To show the importance of **not** using padding when learning keypoints, we provide two tables (cf. Tab. 4) as evidence of the adverse effects of using it.

3. Implementation details

3.1. Data augmentation

Here we detail the data augmentation used by SiLK during training, provided by Albumentation library (Fig. 5):

References

[1] 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. 2

[2] David G Lowe. Distinctive image features from scale-invariant keypoints. International journal of computer vision,

HPatches										
	Repeatability		Hom. Est. Acc.		Hom. Est. AUC		MMA		# of keypoints	
k	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	pre-match	post-match
1.0k	0.43	0.61	0.53	0.81	0.32	0.58	0.52	0.63	1000	389
2.5k	0.50	0.70	0.58	0.83	0.37	0.62	0.55	0.67	2500	1006
5.0k	0.56	0.76	0.60	0.85	0.39	0.64	0.57	0.69	5000	2074
7.5k	0.59	0.79	0.60	0.85	0.39	0.64	0.58	0.70	7500	3169
10.0k	0.62	0.81	0.62	0.87	0.40	0.66	0.59	0.71	10000	4283
12.5k	0.64	0.82	0.63	0.87	0.40	0.66	0.59	0.72	12500	5408
15.0k	0.66	0.83	0.62	0.86	0.40	0.66	0.60	0.72	15000	6541
17.5k	0.67	0.84	0.62	0.86	0.41	0.66	0.60	0.73	17500	7677
20.0k	0.69	0.85	0.63	0.86	0.41	0.66	0.60	0.73	20000	8821
22.5k	0.70	0.85	0.62	0.87	0.41	0.66	0.60	0.73	22500	9963
25.0k	0.71	0.86	0.64	0.87	0.42	0.67	0.60	0.73	25000	11106

Table 2: Numerical results used in Fig. 3.

HPatches									
		Hom. Est. Acc.		Hom. Est. AUC		MMA		# of keypoints	
	Threshold	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 1$	$\epsilon = 3$	pre-match	post-match
ratio-test	1.00	0.62	0.87	0.40	0.66	0.59	0.71	10000	4283
	0.95	0.61	0.87	0.40	0.65	0.62	0.75	10000	3771
	0.90	0.61	0.86	0.40	0.65	0.66	0.79	10000	3314
	0.85	0.59	0.84	0.40	0.64	0.69	0.83	10000	2926
	0.80	0.60	0.83	0.40	0.63	0.72	0.86	10000	2599
	0.75	0.58	0.82	0.40	0.62	0.74	0.88	10000	2314
	0.70	0.56	0.82	0.39	0.61	0.76	0.89	10000	2061
	0.65	0.54	0.78	0.39	0.59	0.77	0.90	10000	1830
	0.60	0.53	0.78	0.38	0.58	0.77	0.91	10000	1620
	0.55	0.51	0.75	0.37	0.57	0.77	0.91	10000	1427
	0.50	0.49	0.76	0.37	0.55	0.76	0.90	10000	1248
double-softmax	1.00	0.62	0.86	0.40	0.65	0.52	0.63	10000	5321
	0.95	0.59	0.86	0.39	0.64	0.68	0.83	10000	3650
	0.90	0.59	0.84	0.39	0.63	0.73	0.88	10000	3244
	0.85	0.55	0.83	0.39	0.62	0.75	0.91	10000	2923
	0.80	0.56	0.81	0.39	0.61	0.77	0.92	10000	2616
	0.75	0.56	0.80	0.39	0.61	0.77	0.92	10000	2319
	0.70	0.55	0.79	0.39	0.60	0.78	0.93	10000	2041
	0.65	0.55	0.79	0.38	0.59	0.78	0.93	10000	1782
	0.60	0.53	0.78	0.38	0.58	0.79	0.93	10000	1547
	0.55	0.53	0.76	0.38	0.57	0.79	0.92	10000	1335
	0.50	0.51	0.76	0.38	0.57	0.79	0.92	10000	1141

Table 3: Numerical results used in Fig. 4.

60(2):91–110, 2004. 3

[3] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2d2: repeatable and reliable detector and descriptor. arXiv preprint arXiv:1906.06195, 2019. 2

[4] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8922–8931, 2021. 2, 3

[5] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. Advances in Neural Information Processing Systems, 33:14254–14265, 2020. 2

	HPatches									
	Repeatability		Hom. Est. Acc.		Hom. Est. AUC		MMA		# of keypoints	
padding	0.59	0.79	0.59	0.84	0.39	0.63	0.57	0.68	10000	4222
no padding	**0.62**	**0.81**	**0.62**	**0.87**	**0.40**	**0.66**	**0.59**	**0.71**	10000	4283

	ScanNet														
	Rotation					Translation					Chamfer				
	Accuracy ↑			Error ↓		Accuracy ↑			Error ↓		Accuracy ↑			Error ↓	
	5°	10°	45°	Mean	Med.	5	10	25	Mean	Med.	1	5	10	Mean	Med.
padding	93.7	97.2	**99.7**	2.1	0.9	75.2	89.8	97.4	5.2	2.5	85.6	95.9	97.8	4.6	**0.1**
no padding	**98.1**	**99.0**	99.6	**1.7**	**0.8**	**82.9**	**94.8**	**99.0**	**4.1**	**2.1**	**92.8**	**98.3**	**99.1**	**4.3**	**0.1**

Table 4: **Avoid using padding to learn good keypoints.**

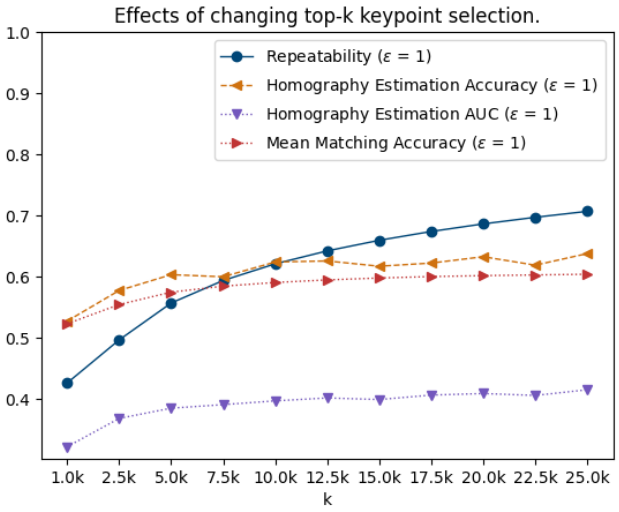

Figure 3: **Increasing top-k keypoint selection gives an initial boost in performance, but tend to get dimishing returns for** $k > 10,000$**.**

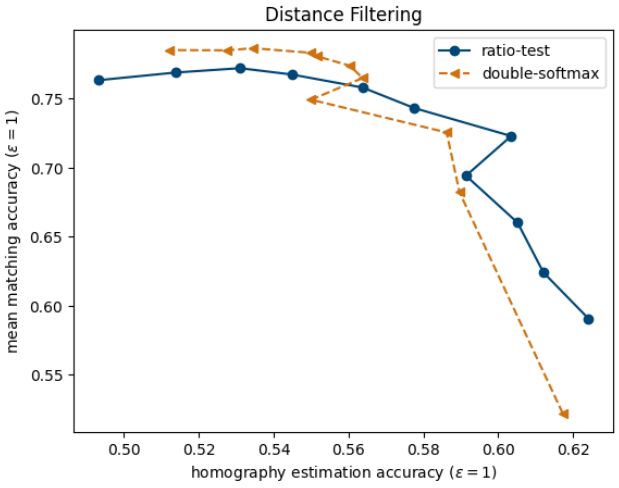

Figure 4: **MMA / Homography trade-off can be controlled with ratio-test and double-softmax filtering.** Different threshold values are tested (between 0.5 and 1) using both methods.

```
import albumentations as A

silk_augmentation = A.Compose([
    A.RandomGamma(
        p=0.1, gamma_limit=(15, 65)
    ),
    A.HueSaturationValue(
        p=0.1, val_shift_limit=(-100, -40)
    ),
    A.Blur(
        p=0.1, blur_limit=(3, 9)
    ),
    A.MotionBlur(
        p=0.2, blur_limit=(3, 25)
    ),
    A.RandomBrightnessContrast(
        p=0.5,
        brightness_limit=(-0.3, 0.0),
        contrast_limit=(-0.5, 0.3)
    ),
    A.GaussNoise(p=0.5),
], p=0.95)
```

Figure 5: Pseudo-code: Data augmentation for SiLK.