SuperGlue: Learning Feature Matching with Graph Neural Networks

Paul-Edouard Sarlin[1] Daniel DeTone[2] Tomasz Malisiewicz[2] Andrew Rabinovich[2]

Appendix

A. Detailed results

A.1. Homography estimation

Qualitative results: A full page of qualitative results of SuperGlue matching on synthetic and real homographies can be seen in Figure 6.

Synthetic dataset: We take a more detailed look at the homography evaluation from Section 5.1. Figure 1 shows the match precision at several correctness pixel thresholds and the cumulative error curve of homography estimation. SuperGlue dominates across all pixel correctness thresholds.

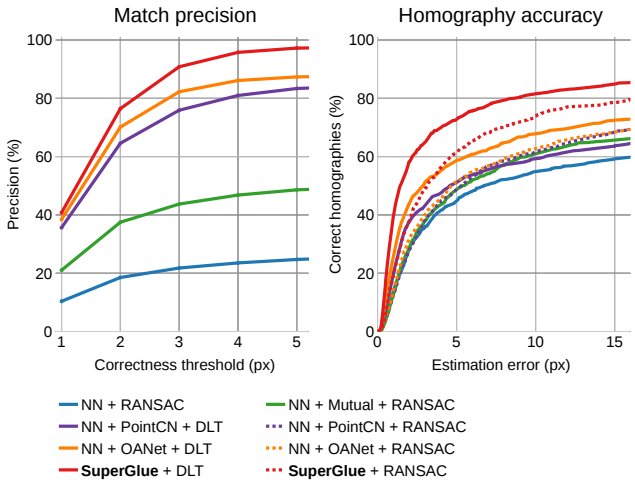

Figure 1: **Details of the homography evaluation.** SuperGlue exhibits higher precision and homography accuracy at all thresholds. High precision results in more accurate estimation with DLT than with RANSAC.

Local features	Matcher	Viewpoint		Illumination	
		P	R	P	R
SuperPoint	NN	39.7	81.7	51.1	84.9
	NN + mutual	65.6	77.1	74.2	80.7
	NN + PointCN	87.6	80.7	94.5	82.6
	NN + OANet	90.4	81.2	**96.3**	83.5
	SuperGlue	**91.4**	**95.7**	89.1	**91.7**

Table 1: **Generalization to real data.** We show the precision (P) and recall (R) of the methods trained on our synthetic homography dataset (see Section 5.1) on the viewpoint and illumination subsets of the HPatches dataset. While trained on synthetic homographies, SuperGlue generalizes well to real data.

[1] ETH Zurich
[2] Magic Leap, Inc.

HPatches: We assess the generalization ability of SuperGlue on real data with the HPatches [2] dataset, as done in previous works [5, 11]. This dataset depicts planar scenes with ground truth homographies and contains 295 image pairs with viewpoint changes and 285 pairs with illumination changes. We evaluate the models trained on the synthetic dataset (see Section 5.1). The HPatches experiment is summarized in Table 1. As previously observed in the synthetic homography experiments, SuperGlue has significantly higher recall than all matchers relying on the NN search. We attribute the remaining gap in recall to several challenging pairs for which SuperPoint does not detect enough repeatable keypoints. Nevertheless, synthetic-dataset trained SuperGlue generalizes well to real data.

A.2. Indoor pose estimation

Qualitative results: More visualizations of matches computed by SuperGlue on indoor images are shown in Figure 7, and highlight the extreme difficulty of the wide-baseline image pairs that constitute our evaluation dataset.

ScanNet: We present more details regarding the results on ScanNet (Section 5.2), only analyzing the methods which use SuperPoint local features. Figure 2 plots the cumulative pose estimation error curve and the trade-off between precision and number of correct matches. We compute the correctness from the reprojection error (using the ground truth depth and a threshold of 10 pixels), and, for keypoints with invalid depth, from the symmetric epipolar error. We obtain curves by varying the confidence thresholds of PointCN, OANet, and SuperGlue. At evaluation, we use the original value 0.5 for the former two, and 0.2 for SuperGlue.

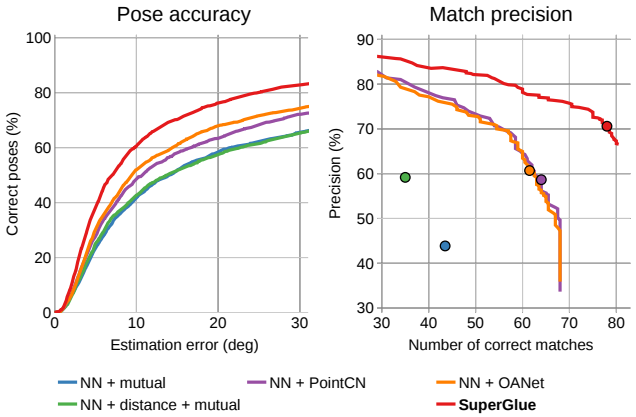

Figure 2: **Details of the ScanNet evaluation.** Poses estimated with SuperGlue are more accurate at all error thresholds. SuperGlue offers the best trade-off between precision and number of correct matches, which are both critical for accurate and robust pose estimation.

Local features	Matcher	Exact AUC			Approx. AUC [19]		
		5°	10°	20°	5°	10°	20°
ContextDesc	NN + ratio test	26.09	45.52	63.07	53.00	63.13	73.00
SIFT	NN + ratio test	24.09	40.71	58.14	45.12	55.81	67.20
	NN + OANet*	28.76	48.42	66.18	55.50	65.94	76.17
	NN + OANet	29.15	48.12	65.08	55.06	64.97	74.83
	SuperGlue	**30.49**	**51.29**	**69.72**	**59.25**	**70.38**	**80.44**
SuperPoint	NN + mutual	16.94	30.39	45.72	35.00	43.12	54.05
	NN + OANet	26.82	45.04	62.17	50.94	61.41	71.77
	SuperGlue	**38.72**	**59.13**	**75.81**	**67.75**	**77.41**	**85.70**

Table 2: **Outdoor pose estimation on YFCC100M pairs.** The evaluation is performed on the same image pairs as in OANet [19] using both their approximate and our exact AUC. SuperGlue consistently improves over the baselines when using either SIFT and SuperPoint.

A.3. Outdoor pose estimation

Qualitative results: Figure 8 shows additional results on the Phototourism test set and the MegaDepth validation set.

YFCC100M: While the PhotoTourism [1] and Zhang *et al.*'s [19] test sets are both based on YFCC100M [16], they use different scenes and pairs. For the sake of comparability, we also evaluate SuperGlue on the same evaluation pairs as in OANet [19], using their evaluation metrics. We include an OANet model (*) retrained on their training set (instead of MegaDepth) using root-normalized SIFT. The results are shown in Table 2.

As observed in Section 5.3 when evaluating on the PhotoTourism dataset, SuperGlue consistently improves over all baselines for both SIFT and SuperPoint. For SIFT, the improvement over OANet is decreased, which we attribute to the significantly higher overlap and lower difficulty of the pairs used by [19]. While the approximate AUC tends to overestimate the accuracy, it results in an identical ranking of the methods. The numbers for OANet with SIFT and SuperPoint are consistent with the ones reported in their paper.

Method	Correctly localized queries (%)			# features
	.5m/2°	1m/5°	5m/10°	
R2D2 [11]	**46.9**	66.3	**88.8**	20k
D2-Net [6]	45.9	68.4	**88.8**	15k
UR2KID [18]	**46.9**	67.3	**88.8**	15k
SuperPoint+NN+mutual	43.9	59.2	76.5	4k
SuperPoint+SuperGlue	45.9	**70.4**	**88.8**	4k

Table 3: **Visual localization on Aachen Day-Night.** SuperGlue significantly improves the performance of SuperPoint for localization, reaching new state-of-the-art results with comparably fewer keypoints.

B. SuperGlue for visual localization

Visual localization: While two-view relative pose estimation is an important fundamental problem, advances in image matching can directly benefit practical tasks like visual localization [13, 12], which aims at estimating the absolute pose of a query image with respect to a 3D model. Moreover, real-world localization scenarios exhibit significantly higher scene diversity and more challenging conditions, such as larger viewpoint and illumination changes, than phototourism datasets of popular landmarks.

Evaluation: The Aachen Day-Night benchmark [14, 13] evaluates local feature matching for day-night localization. We extract up to 4096 keypoints per images with SuperPoint, match them with SuperGlue, triangulate an SfM model from posed day-time database images, and register night-time query images with the 2D-2D matches and COLMAP [15]. The evaluation server[1] computes the percentage of queries localized within several distance and orientation thresholds. As reported in Table 3, SuperPoint+SuperGlue performs similarly or better than all existing approaches despite using significantly fewer keypoints. Figure 3 shows challenging day-night image pairs.

[1] https://www.visuallocalization.net/

Figure 3: **Matching challenging day-night pairs with SuperGlue.** We show predicted correspondences between night-time queries and day-time databases images of the Aachen Day-Night dataset. The correspondences are colored as RANSAC inliers in green or outliers in red. Although the outdoor training set has few night images, SuperGlue generalizes well to such extreme illumination changes. Moreover, it can accurately match building facades with repeated patterns like windows.

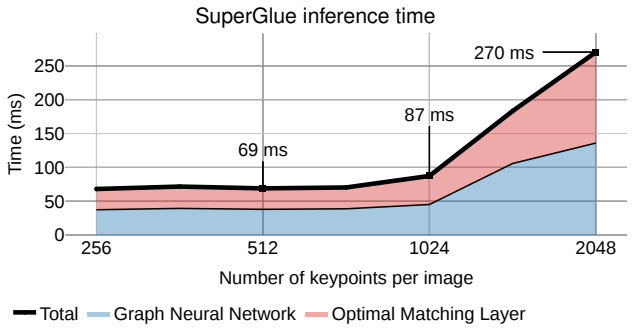

Figure 4: **SuperGlue detailed inference time.** SuperGlue's two main blocks, the Graph Neural Network and the Optimal Matching Layer, have similar computational costs. For 512 and 1024 keypoints per image, SuperGlue runs at 14.5 and 11.5 FPS, respectively.

C. Timing and model parameters

Timing: We measure the run-time of SuperGlue and its two major blocks, the Graph Neural Network and the Optimal Matching Layer, for different numbers of keypoints per image. The measurements are performed on an NVIDIA GeForce GTX 1080 GPU across 500 runs. See Figure 4.

Model Parameters: The Keypoint Encoder MLP has 5 layers, mapping positions to dimensions of size $(32, 64, 128, 256, D)$, yielding 100k parameters. Each layer has the three projection matrices, and an extra $\mathbf{W}^O$ to deal with the multi-head output. The message update MLP has 2 layers and maps to dimensions $(2D, D)$. Both MLPs use BatchNorm and ReLUs. Each layer has 0.66M parameters. SuperGlue has 18 layers, with a total of 12M parameters.

D. Analyzing attention

Quantitative analysis: We compute the spatial extent of the attention weights – the *attention span* – for all layers and all keypoints. The self-attention span corresponds to the distance in pixel space between one keypoint i and all the others j, weighted by the attention weight α_{ij}, and averaged for all queries. The cross-attention span corresponds to the average distance between the final predicted match and all the attended keypoints j. We average the spans over 100 ScanNet pairs and plot in Figure 5 the minimum across all heads for each layer, with 95% confidence intervals.

The spans of both self- and cross-attention tend to decrease throughout the layers, by more than a factor of 10 between the first and the last layer. SuperGlue initially attends to keypoints covering a large area of the image, and later focuses on specific locations – the self-attention attends to a small neighborhood around the keypoint, while the cross-attention narrows its search to the vicinity of the true match. Intermediate layers have oscillating spans, hinting at a more complex process.

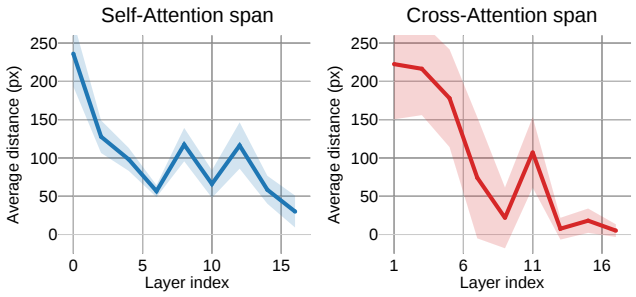

Figure 5: **Attention spans throughout SuperGlue.** We plot the attention span, a measure of the attention's spatial dispersion, vs. layer index. For both types of attention, the span tends to decrease deeper in the network as SuperGlue focuses on specific locations. See an example in Figure 9.

Qualitative example: We analyze the attention patterns of a specific example in Figure 9. Our observations are consistent with the attention span trends reported in Figure 5.

E. Experimental details

In this section, we provide details on the training and evaluation of SuperGlue. The trained models and the evaluation code and image pairs are publicly available at `github.com/magicleap/SuperGluePretrainedNetwork`.

Choice of indoor dataset: Previous works on inlier classification [9, 19, 3] evaluate indoor pose estimation on the SUN3D dataset [17]. Camera poses in SUN3D are estimated from SIFT-based sparse SfM, while ScanNet leverages RGB-D fusion and optimization [4], resulting in significantly more accurate poses. This makes ScanNet more suitable for generating accurate correspondence labels and evaluating pose estimation. We additionally noticed that the SUN3D image pairs used by Zhang *et al.* [19] have generally small baseline and rotation angle. This makes the essential matrix estimation degenerate [7] and the angular translation error ill-defined. In contrast, our ScanNet wide-baseline pairs have significantly more diversity in baselines and rotation, and thus do not suffer from the aforementioned issues.

Homography estimation – Section 5.1: The test set contains 1024 pairs of 640×480 images. Homographies are generated by applying random perspective, scaling, rotation, and translation to the original full-sized images, to avoid bordering artifacts. We evaluate with the 512 top-scoring keypoints detected by SuperPoint with a Non-Maximum Suppression (NMS) radius of 4 pixels. Correspondences are deemed correct if they have a reprojection error lower than 3 pixels. We use the OpenCV function `findHomography` with 3000 iterations and a RANSAC inlier threshold of 3 pixels.

Indoor pose estimation – Section 5.2: The overlap score between two images A and B is the average ratio of pixels in A that are visible in B (and vice versa), after accounting for missing depth values and occlusion (by checking for consistency in the depth). We train and evaluate with pairs that have an overlap score in $[0.4, 0.8]$. For training, we sample at each epoch 200 pairs per scene, similarly as in [6]. The test set is generated by subsampling the sequences by 15 and subsequently randomly sampling 15 pairs for each of the 300 sequences. We resize all ScanNet images and depth maps to 640×480. We detect up to 1024 SuperPoint keypoints (using the publicly available trained model[2] with NMS radius of 4) and 2048 SIFT keypoints (using OpenCV's implementation). Poses are computed by first estimating the essential matrix with OpenCV's `findEssentialMat` and RANSAC with an inlier threshold of 1 pixel divided by the focal length, followed by `recoverPose`. In contrast with previous works [9, 19, 3], we compute a more accurate AUC using explicit integration rather than coarse histograms. The precision (P) is the average ratio of the number of correct matches over the total number of estimated matches. The matching score (MS) is the average ratio of the number of correct matches over the total number of detected keypoints. It does not account for the pair overlap and decreases with the number of covisible keypoints. A match is deemed correct if its epipolar distance is lower than $5 \cdot 10^{-4}$.

Outdoor pose estimation – Section 5.3: For training on Megadepth, the overlap score is the ratio of triangulated keypoints that are visible in the two images, as in [6]. We sample pairs with an overlap score in $[0.1, 0.7]$ at each epoch. We evaluate on all 11 scenes of the PhotoTourism dataset and reuse the overlap score based on bounding boxes computed by Ono *et al*. [10], with a selection range of $[0.1, 0.4]$. Images are resized so that their longest dimension is equal to 1600 pixels and rotated upright using their EXIF data. We detect 2048 keypoints for both SIFT and SuperPoint (with an NMS radius of 3). The epipolar correctness threshold is here 10^{-4}. Other evaluation parameters are identical to the ones used for the indoor evaluation.

Training of SuperGlue: For training on homography/indoor/outdoor data, we use the Adam optimizer [8] with a constant leaning rate of 10^{-4} for the first 200k/100k/50k iterations, followed by an exponential decay of 0.999998/0.999992/0.999992 until iteration 900k. When using SuperPoint features, we employ batches with 32/64/16 image pairs and a fixed number of 512/400/1024 keypoints per image. For SIFT features we use 1024 keypoints and 24 pairs. Due to the limited number of training scenes, the outdoor model weights are initialized with the homography model weights. Before the keypoint encoder, the keypoints are normalized by the largest dimension of the image.

Ground truth correspondences $\mathcal{M}$ and unmatched sets $\mathcal{I}$ and $\mathcal{J}$ are generated by first computing the $M \times N$ reprojection matrix between all detected keypoints using the ground truth homography or pose and depth. Correspondences are entries with a reprojection error that is a minimum along both rows and columns, and that is lower than a given threshold: 3, 5, and 3 pixels for homographies, indoor, and outdoor matching respectively. For homographies, unmatched keypoints are simply the ones that do not appear in $\mathcal{M}$. For indoor and outdoor matching, because of errors in the pose and depth, unmatched keypoints must additionally have a minimum reprojection error larger than 15 and 5 pixels, respectively. This allows us to ignore labels for keypoints whose correspondences are ambiguous, while still providing some supervision through the normalization induced by the Sinkhorn algorithm.

Ablation study – Section 5.4: The "No Graph Neural Net" baseline replaces the Graph Neural Network with a single linear projection, but retains the Keypoint Encoder and the Optimal Matching Layer. The "No cross-attention" baseline replace all cross-attention layers by self-attention: it has the same number of parameters as the full model, and acts like a Siamese network. The "No positional encoding" baseline simply removes the Keypoint Encoder and only uses the visual descriptors as input.

End-to-end training – Section 5.4: Two copies of SuperPoint, for detection and description, are initialized with the original weights. The detection network is frozen and gradients are propagated through the descriptor network only, flowing from SuperGlue - no additional losses are used.

[2] github.com/magicleap/SuperPointPretrainedNetwork

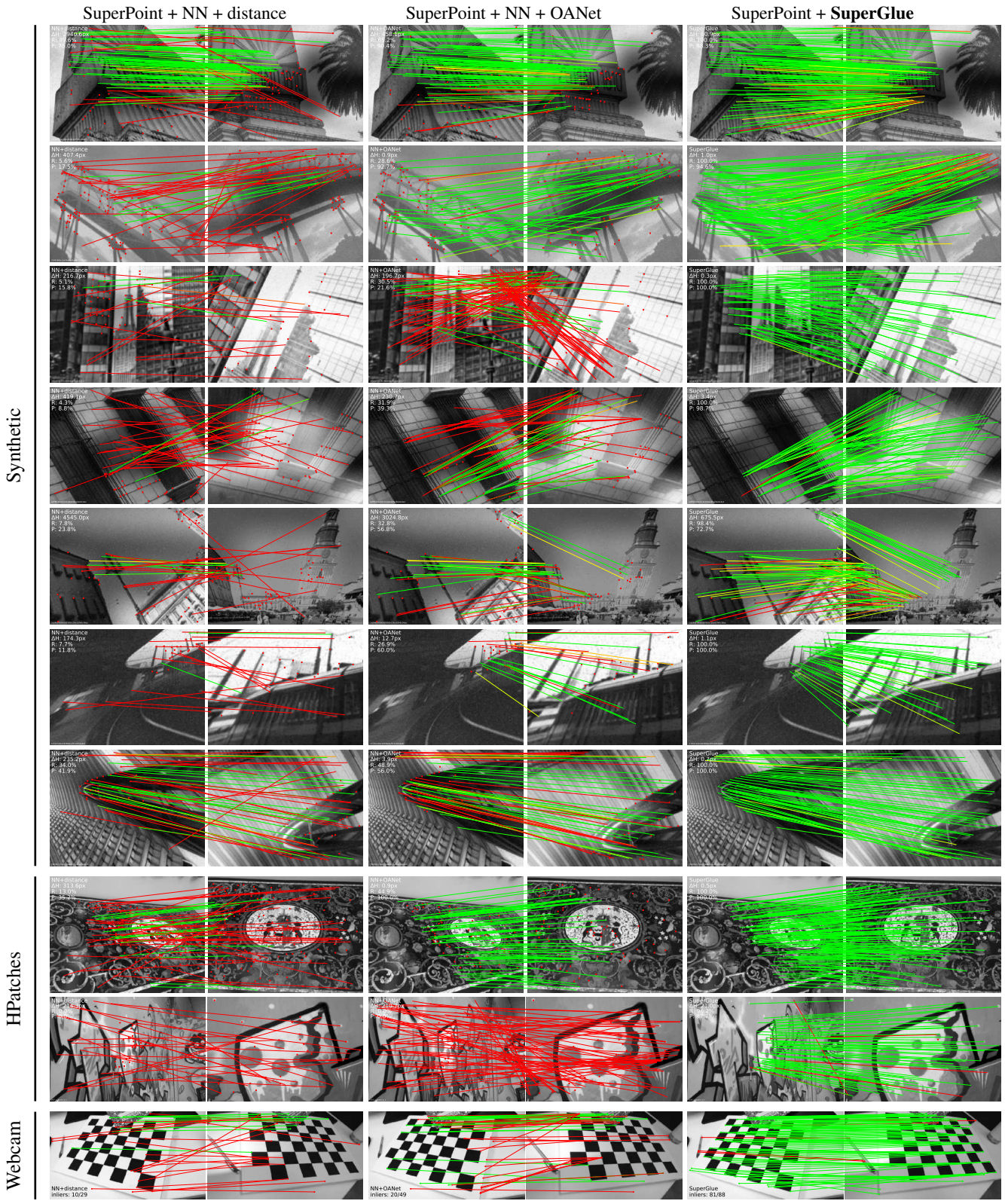

Figure 6: **More homography examples.** We show point correspondences on our synthetic dataset (see Section 5.1), on real image pairs from HPatches (see Appendix A.1), and a checkerboard image captured by a webcam. SuperGlue consistently estimates more correct matches (green lines) and fewer mismatches (red lines), successfully coping with repeated texture, large viewpoint, and illumination changes.

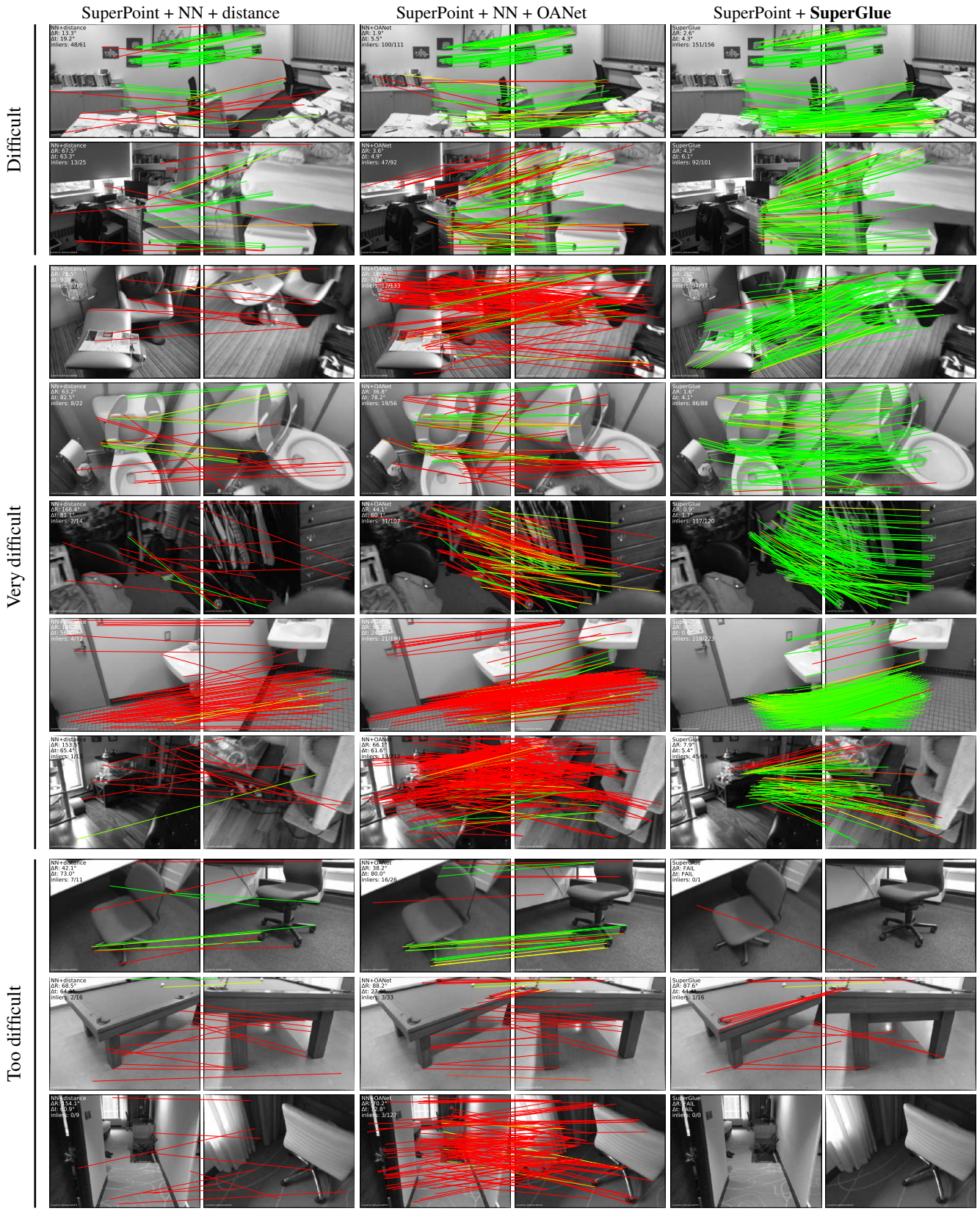

Figure 7: **More indoor examples.** We show both **Difficult** and **Very Difficult** ScanNet indoor examples for which SuperGlue works well, and three **Too Difficult** examples where it fails, either due to unlikely motion or lack of repeatable keypoints (last two rows). Correct matches are green lines and mismatches are red lines. See details in Section 5.2.

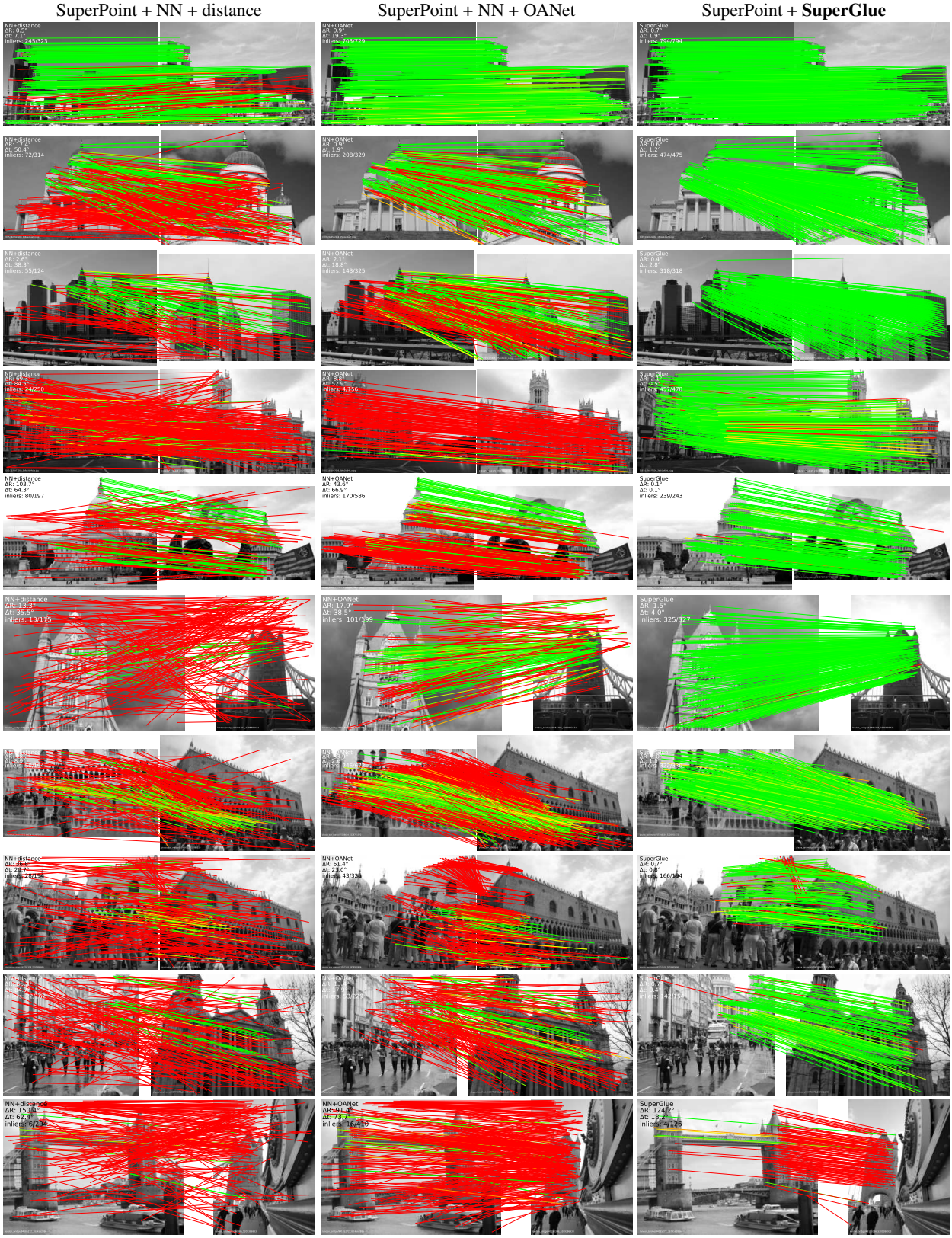

Figure 8: **More outdoor examples.** We show results on the MegaDepth validation and the PhotoTourism test sets. Correct matches are green lines and mismatches are red lines. The last row shows a failure case, where SuperGlue focuses on the incorrect self-similarity. See details in Section 5.3.

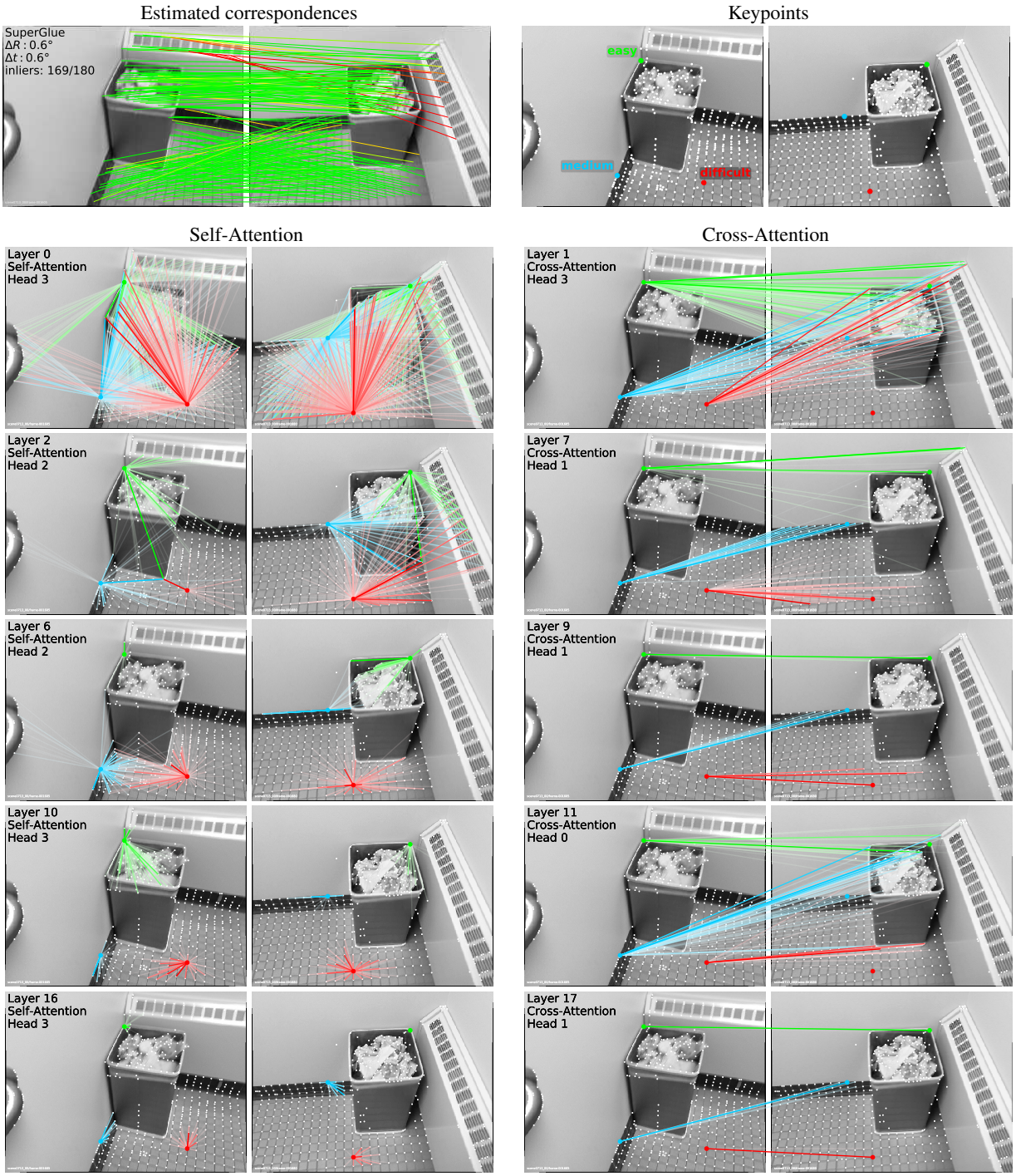

Figure 9: **Attention patterns across layers.** For this image pair (correctly matched by SuperGlue), we look at three specific keypoints that can be matched with different levels of difficulty: the easy keypoint, the medium keypoint, and the difficult keypoint. We visualize self- and cross-attention weights (within images A and B, and from A to B, respectively) of selected layers and heads, varying the edge opacity with α_{ij}. The self-attention initially attends all over the image (row 1), and gradually focuses on a small neighborhood around each keypoint (last row). Similarly, some cross-attention heads focus on candidate matches, and successively reduce the set that is inspected. The easy keypoint is matched as early as layer 9, while more difficult ones are only matched at the last layer. Similarly as in Figure 5, the self- and cross-attention spans generally shrink throughout the layers. They however increase in layer 11, which attends to other locations – seemingly distinctive ones – that are further away. We hypothesize that SuperGlue attempts to disambiguate challenging matches using additional context.

References

[1] Phototourism Challenge, CVPR 2019 Image Matching Workshop. https://image-matching-workshop.github.io. Accessed November 8, 2019. 2

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 1

[3] Eric Brachmann and Carsten Rother. Neural-Guided RANSAC: Learning where to sample model hypotheses. In *ICCV*, 2019. 3, 4

[4] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(3):24, 2017. 3

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *CVPR Workshop on Deep Learning for Visual SLAM*, 2018. 1

[6] 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. In *CVPR*, 2019. 2, 4

[7] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 4

[9] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *CVPR*, 2018. 3, 4

[10] Yuki Ono, Eduard Trulls, Pascal Fua, and Kwang Moo Yi. LF-Net: Learning local features from images. In *NeurIPS*, 2018. 4

[11] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. R2D2: Repeatable and reliable detector and descriptor. In *NeurIPS*, 2019. 1, 2

[12] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 2

[13] Torsten Sattler, Will Maddern, Carl Toft, Akihiko Torii, Lars Hammarstrand, Erik Stenborg, Daniel Safari, Masatoshi Okutomi, Marc Pollefeys, Josef Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *CVPR*, 2018. 2

[14] Torsten Sattler, Tobias Weyand, Bastian Leibe, and Leif Kobbelt. Image retrieval for image-based localization revisited. In *BMVC*, 2012. 2

[15] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[16] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 2

[17] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *ICCV*, 2013. 3

[18] Tsun-Yi Yang, Duy-Kien Nguyen, Huub Heijnen, and Vassileios Balntas. UR2KiD: Unifying retrieval, keypoint detection, and keypoint description without local correspondence supervision. *arXiv:2001.07252*, 2020. 2

[19] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *ICCV*, 2019. 2, 3, 4