SuperGlue: Learning Feature Matching with Graph Neural Networks

Paul-Edouard Sarlin Daniel DeTone Tomasz Malisiewicz Andrew Rabinovich

ETH Zurich, Magic Leap, Inc.

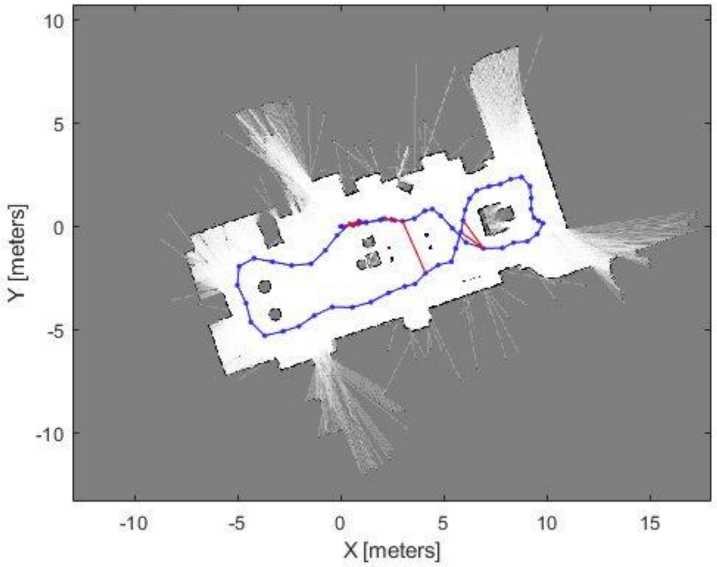
CVPR 2020 (oral)

Presenter: Minho Park

Task

- Structure from Motion (SfM): https://www.youtube.com/watch?v=i7ierVkXYa8
- Simultaneous Localization and Mapping (SLAM): https://kr.mathworks.com/discovery/slam.html

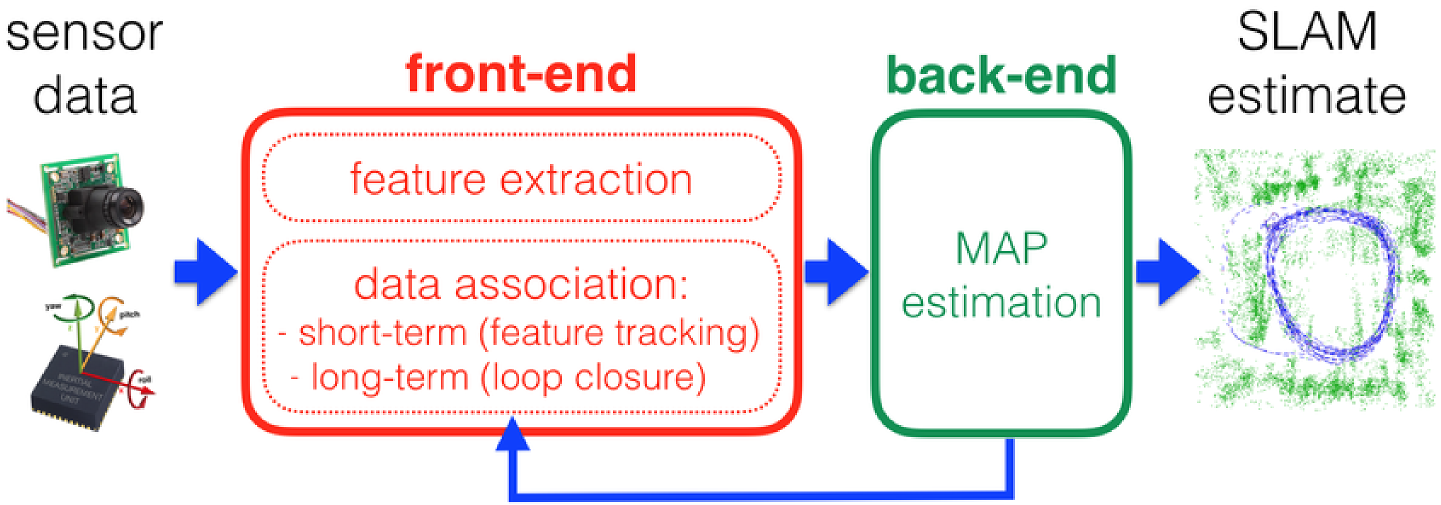
Reconstructed 3D Model from SfM



Occupancy grid map built using Lidar SLAM

SLAM System

- **Front-end**: Visual feature extraction
- **Back-end**: Bundle adjustment or pose estimation
- **Middle-end**: Learnable SuperGlue



sensor data

front-end

feature extraction

data association:
- short-term (feature tracking)
- long-term (loop closure)

back-end

MAP estimation

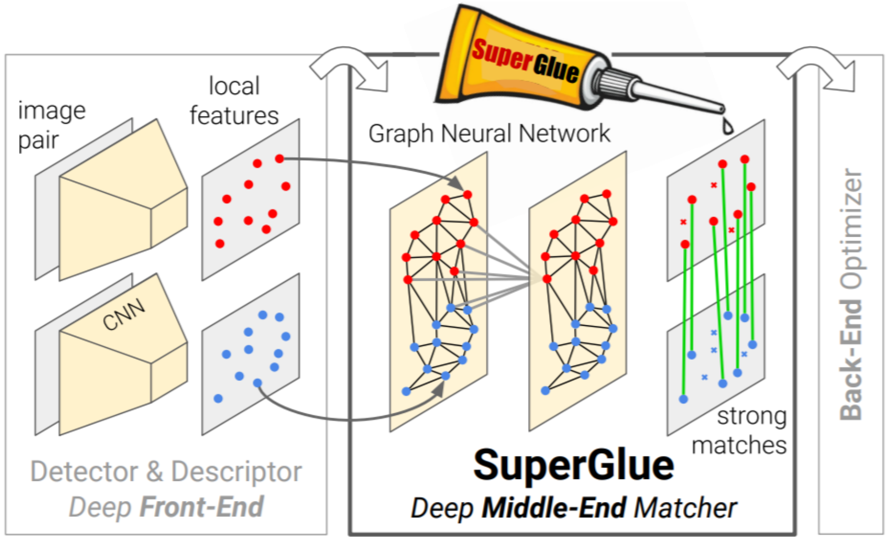
SLAM estimate

Fig 2. Front end and back end in a typical SLAM system. The back end can provide feedback to the front end for loop closure detection and verification.

Cadena, Cesar, et al. "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age." *IEEE Transactions on robotics* 32.6 (2016): 1309-1332.

SuperGlue

- **Middle-end**: Learnable SuperGlue



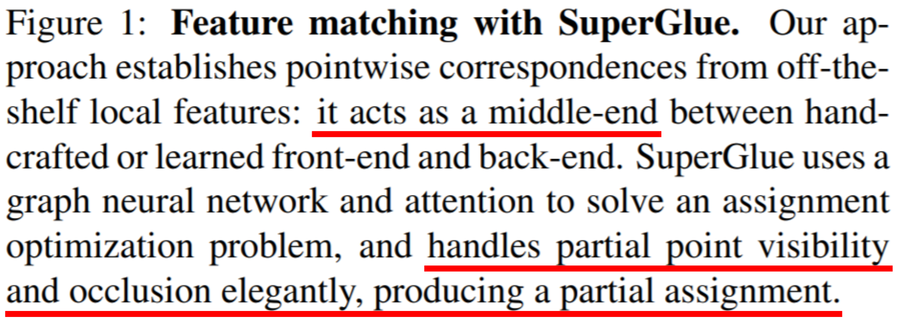


Figure 1: **Feature matching with SuperGlue.** Our approach establishes pointwise correspondences from off-the-shelf local features: it acts as a middle-end between hand-crafted or learned front-end and back-end. SuperGlue uses a graph neural network and attention to solve an assignment optimization problem, and handles partial point visibility and occlusion elegantly, producing a partial assignment.

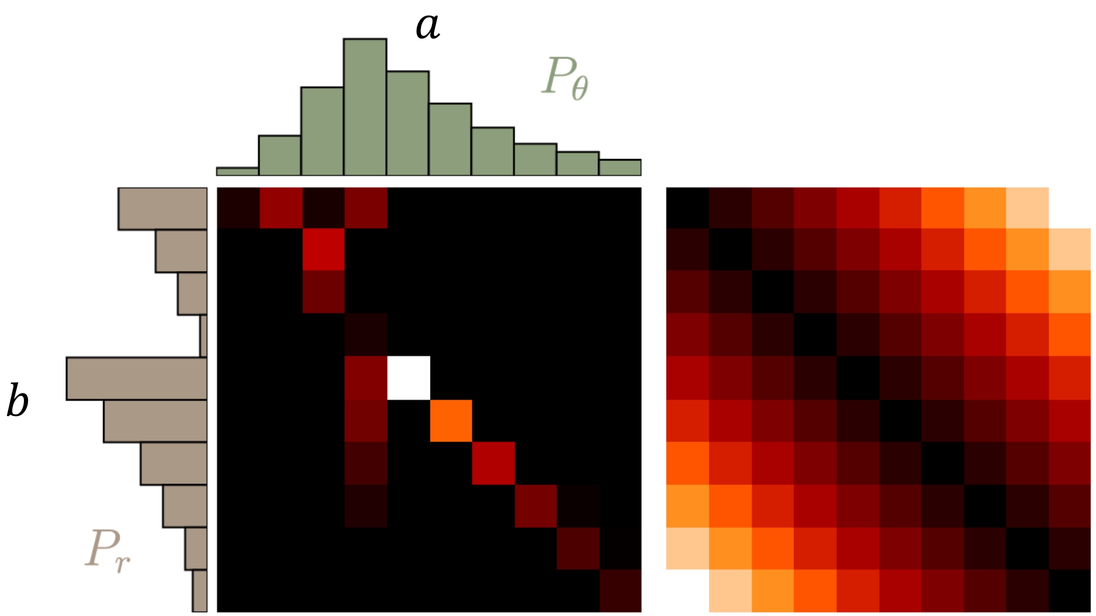
Local Feature Matching

- Local feature matching is based on SIFT.
    1. Detecting interest points.
    2. Computing visual descriptors.
    3. Matching these with a Nearest Neighbor (NN) search.
    4. Filtering incorrect matches.
    5. Estimating a geometric transformation.
- Recent works on deep learning for matching often focus on learning better sparse detectors and local descriptors.

SIFT blog post (KR): https://bskyvision.com/21

Graph Matching

- Graph matching can be represented as the problem of optimal transport.
- It is linear programming that can be solved(optimized) by the Sinkhorn algorithm.



Transportation Plan P Transport cost C

1D optimal transport example
(Euclidean distance case)

$$\min_P \sum_i \sum_j P_{ij} C_{ij}$$

$$s.t. \sum_j P_{ij} = a_i, \qquad \sum_i P_{ij} = b_j$$

$$P_{ij} \geq 0 \quad \forall i, j$$

⇒ Linear Programming

⇒ Can be optimized using Sinkhorn algorithm

Vincent Herrmann blog post, Wasserstein GAN, https://vincentherrmann.github.io/blog/wasserstein/

Motivation of Architecture

- **Regularities of the world:**
- All correspondences for a given image pair derive from a single **epipolar** transform.
- 2D keypoints are usually projections of salient 3D points, like corner or blobs.

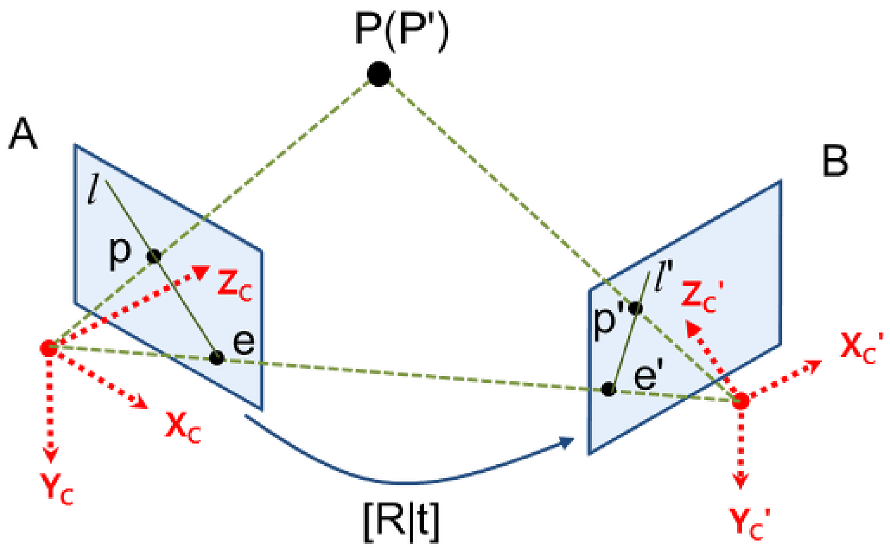


Fig1. Epipole transformation

Darkprogrammer blog post (KR), Geometry #7 Epipolar geometry, https://darkpgmr.tistory.com/83

Motivation of Architecture

- **Physical constraints:**

1. A keypoint can have at most a single correspondence in the other image.
2. Some keypoints will be unmatched due to occlusion and failure of the detector.

Formulation

- $(\boldsymbol{p}, \boldsymbol{d})$: Local features (keypoint position, visual descriptors)
- $\boldsymbol{p}_i = (x, y, c)_i$: x, y are image coordinates, and c is detection confidence
- $\boldsymbol{d}_i \in \mathbb{R}^D$: extracted by a CNN like SuperPoint or traditional descriptors like SIFT
- Image A and B have M and N local features, respectively.
- **Partial assignment (Our goal):**
- $\mathbf{P}\mathbf{1}_N \leq \mathbf{1}_M$, and $\mathbf{P}^T\mathbf{1}_M \leq \mathbf{1}_N$ with partial soft assignment matrix $\mathbf{P} \in [0,1]^{M \times N}$.
- $\begin{bmatrix} & \mathbf{P} & \end{bmatrix} \begin{bmatrix} \mathbf{1}_N \end{bmatrix} \leq \begin{bmatrix} \mathbf{1}_M \end{bmatrix}$

Physical constraints:

1. A keypoint can have at most a single correspon dence in the other image.
2. Some keypoints will be unmatched due to occl usion and failure of the detector.

The SuperGlue Architecture

- **Attentional Graph Neural Network** and **Optimal Matching Layer**

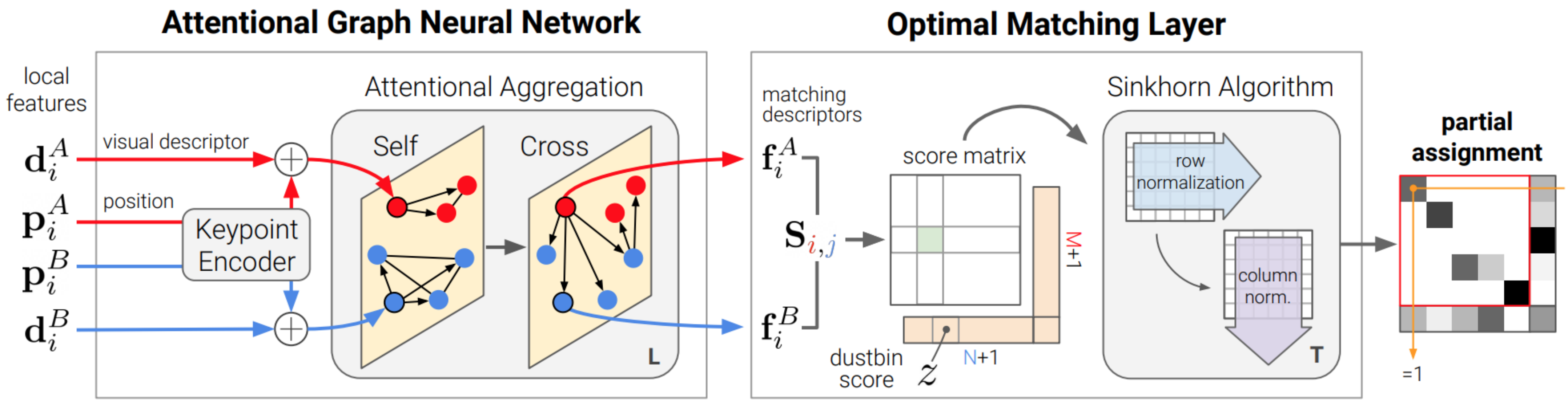
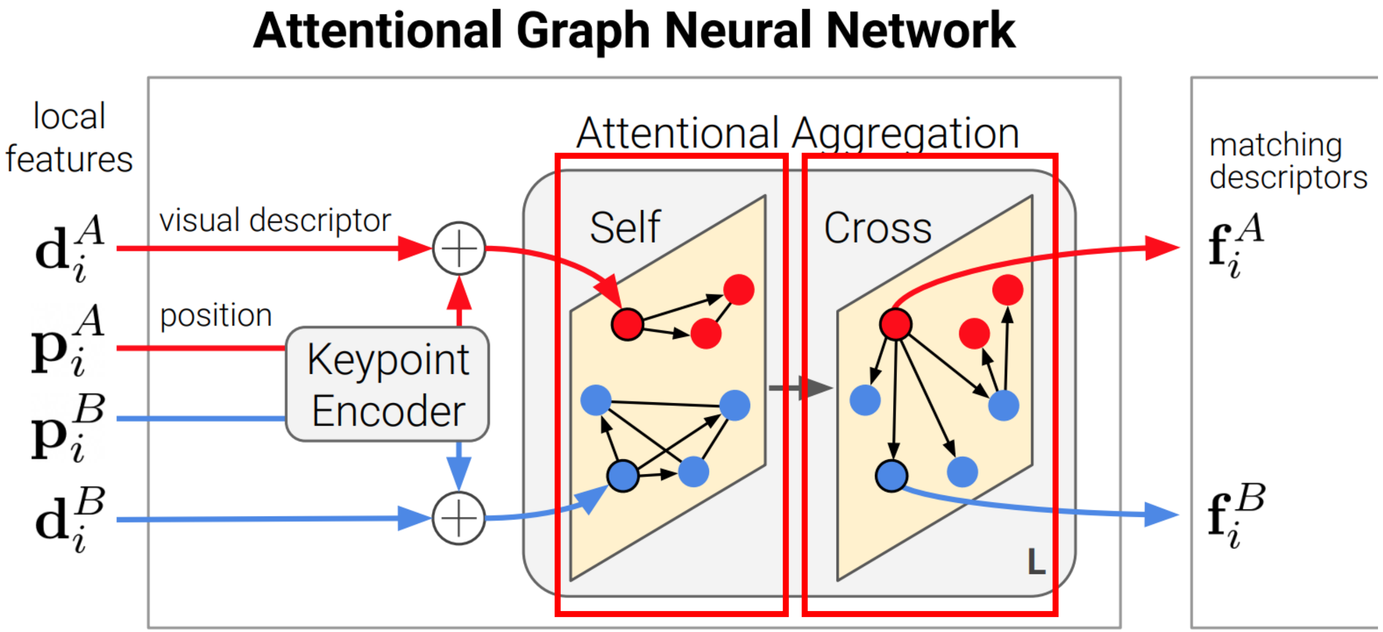


Figure 3: **The SuperGlue architecture.** SuperGlue is made up of two major components: the *attentional graph neural network* (Section 3.1), and the *optimal matching layer* (Section 3.2). The first component uses a *keypoint encoder* to map keypoint positions $\mathbf{p}$ and their visual descriptors $\mathbf{d}$ into a single vector, and then uses alternating self- and cross-attention layers (repeated L times) to create more powerful representations $\mathbf{f}$. The optimal matching layer creates an M by N score matrix, augments it with dustbins, then finds the optimal partial assignment using the Sinkhorn algorithm (for T iterations).

Attentional Graph Neural Network

- We consider its spatial and visual relationship with other co-visible keypoints. **(Self)**
- Knowledge of keypoints in the second image can help to resolve ambiguities. **(Cross)**

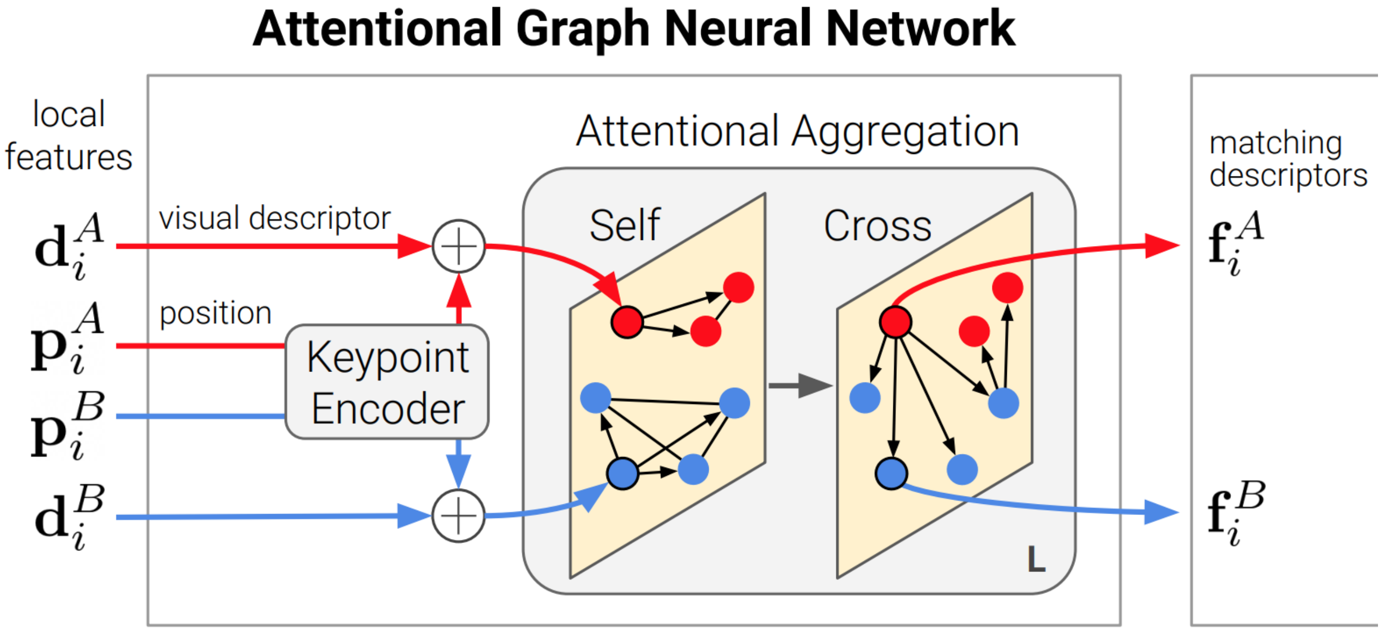


Attentional Graph Neural Network

- **Keypoint Encoder**: Embed the keypoint position into a high dimensional vector with MLP.
  - Work as positional encoding in language processing.

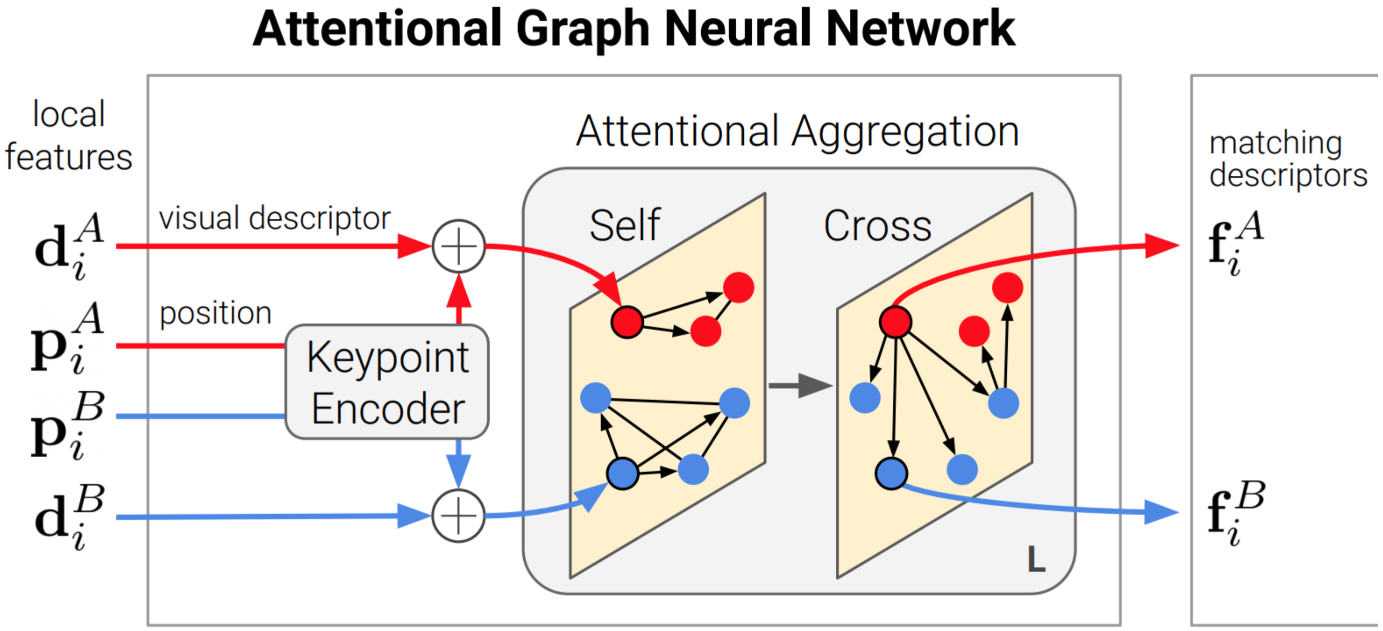
$$\underline{{}^{(0)}x_i} = d_i + \mathrm{MLP}_{\mathrm{enc}}(p_i)$$

Node of GNN



Attentional Graph Neural Network

- **Multiplex Graph Neural Network**:
- Use message passing formulation to propagate information along both type of edges.
  - $\mathcal{E}_{self}, \mathcal{E}_{cross}$: Intra-, inter image edges.



Attentional Graph Neural Network

- **Multiplex Graph Neural Network**:

$$^{(l+1)}x_i^A = {}^{(l)}x_i^A + \mathrm{MLP}\left(\left[{}^{(l)}x_i^A \parallel m_{\mathcal{E}\to i}\right]\right)$$

- $^{(l)}x_i^A$: Intermediate representation for element i in the image A at layer l.
- $m_{\mathcal{E}\to i}$: the result of the aggregation from all keypoints $\{j: (i,j) \in \mathcal{E}\}$, where $\mathcal{E} \in \{\mathcal{E}_{self}, \mathcal{E}_{cross}\}$

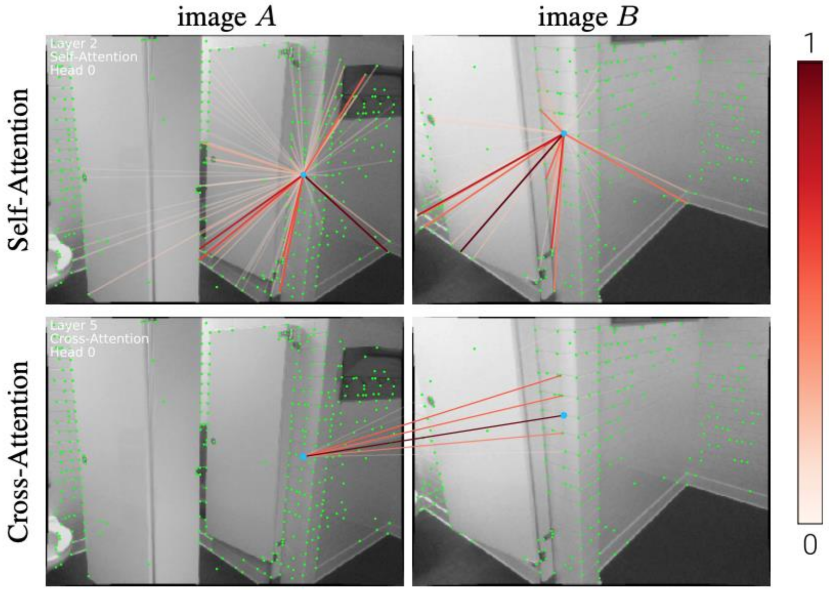
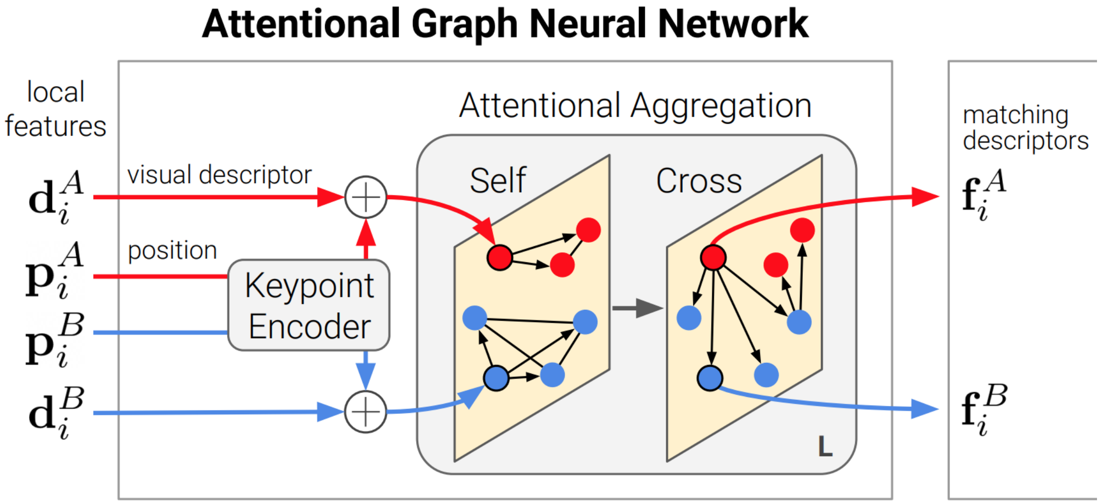


Figure 4: Visualizing self- and cross-attention.



Attentional Graph Neural Network

- **Attentional Aggregation**:

$$ {}^{(l+1)}x_i^A = {}^{(l)}x_i^A + \mathrm{MLP}\left(\left[{}^{(l)}x_i^A \parallel \underline{m_{\mathcal{E}\to i}} \right]\right) $$

- Use multi-head attention to compute the message $m_{\mathcal{E}\to i}$.
- This formulation provides maximum flexibility.

$$ m_{\mathcal{E}\to i} = \sum_{j:(i,j)\in\mathcal{E}} \mathrm{Softmax}_j(q_i^T k_j) v_j $$

$$ q_i = W_1 \, {}^{(l)}x_i^Q + b_1 $$

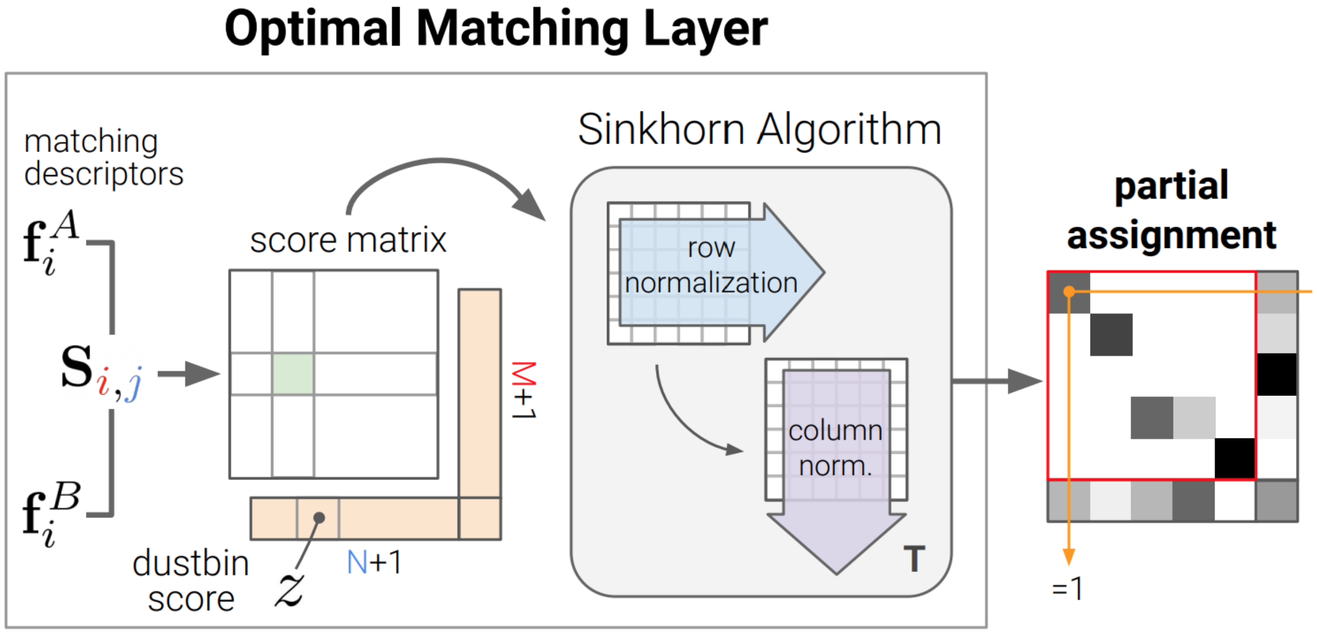
$$ \begin{bmatrix} k_j \\ v_j \end{bmatrix} = \begin{bmatrix} W_2 \\ W_3 \end{bmatrix} {}^{(l)}x_j^S + \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} $$

$$ \underline{f_i^A} = W \, {}^{(L)}x_i^A + b $$

Final matching descriptors

Optimal Matching Layer

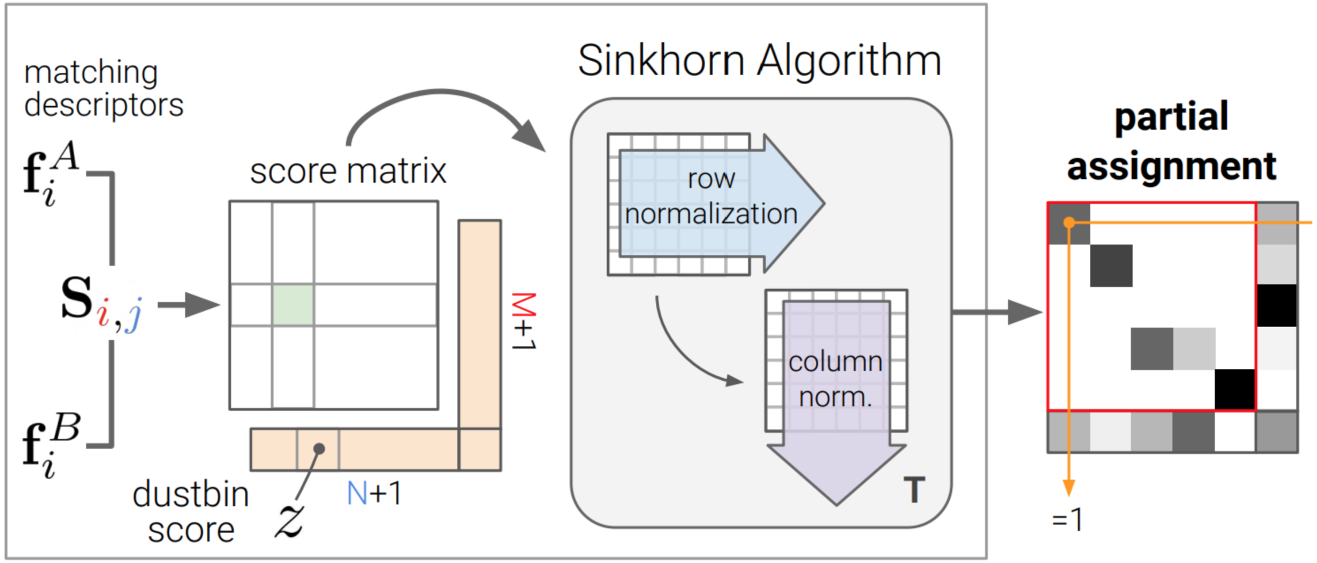
- **Score Prediction**: $S_{i,j} = \langle f_i^A, f_j^B \rangle$
- **Occlusion and Visibility**:
  - Augment each set with a dustbin so that unmatched keypoints are explicitly assigned to it.
  - Dustbin is filled with a single learnable parameter z. (Learnable threshold value)



Optimal Matching Layer

- **Occlusion and Visibility**:
- $\mathbf{P}\mathbf{1}_N \leq \mathbf{1}_M$, and $\mathbf{P}^T\mathbf{1}_M \leq \mathbf{1}_N \rightarrow \bar{\mathbf{P}}\mathbf{1}_{N+1} = \boldsymbol{a}$, and $\bar{\mathbf{P}}^T\mathbf{1}_{M+1} = \boldsymbol{b}$, where $a = \begin{bmatrix}\mathbf{1}_M \\ N\end{bmatrix}$, $\mathrm{b} = \begin{bmatrix}\mathbf{1}_N \\ M\end{bmatrix}$.
  - Imagine ground truth augmented assignment $\bar{\mathbf{P}}$.

Optimal Matching Layer



Sinkhorn Algorithm

- Graph matching can be represented as the problem of optimal transport.
- It is linear programming that can be solved(optimized) by the Sinkhorn algorithm.

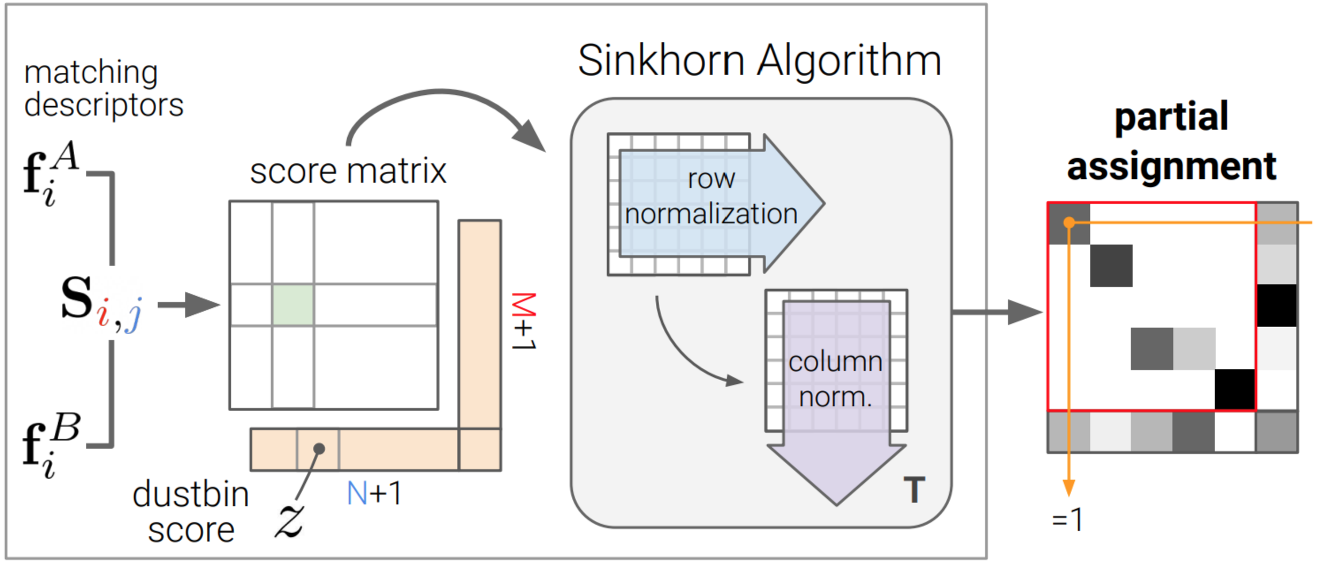
$$
\begin{aligned}
&\boxed{\min_{P}} \sum_i \sum_j P_{ij} C_{ij} \\
&s.t. \sum_j P_{ij} = a_i, \qquad \sum_i P_{ij} = b_j \\
&P_{ij} \geq 0 \quad \forall i, j
\end{aligned}
\qquad \longrightarrow \qquad
\begin{aligned}
&\boxed{\max_{\bar{P}}} \sum_i \sum_j \bar{P}_{ij} \bar{S}_{ij} \\
&s.t. \sum_j \bar{P}_{ij} = a_i, \qquad \sum_i \bar{P}_{ij} = b_j \\
&\bar{P}_{ij} \geq 0 \quad \forall i, j
\end{aligned}
$$

Use negative log optimal transport

Sinkhorn Algorithm

- Iteratively normalizing $\exp \bar{\mathbf{S}}$ along rows and columns, similar to row and column $\mathrm{Softmax}$.
- After T iterations, drop the dustbins: $\mathbf{P} = \bar{\mathbf{P}}_{1:M,1:N}$

Optimal Matching Layer



Loss

- Negative log-likelihood of assignment $\bar{\mathbf{P}}$.

$$\text{Loss} = -\sum_{(i,j)\in\mathcal{M}} \log \bar{\mathbf{P}}_{i,j} - \sum_{i\in\mathcal{I}} \log \bar{\mathbf{P}}_{i,N+1} - \sum_{j\in\mathcal{J}} \log \bar{\mathbf{P}}_{M+1,j}$$

- $\mathcal{M}$: Ground truth matches
- $\mathcal{I}$: Ground truth unmatched labels in image $\mathcal{A}$.
- $\mathcal{J}$: Ground truth unmatched labels in image $\mathcal{B}$.

Wrap Up

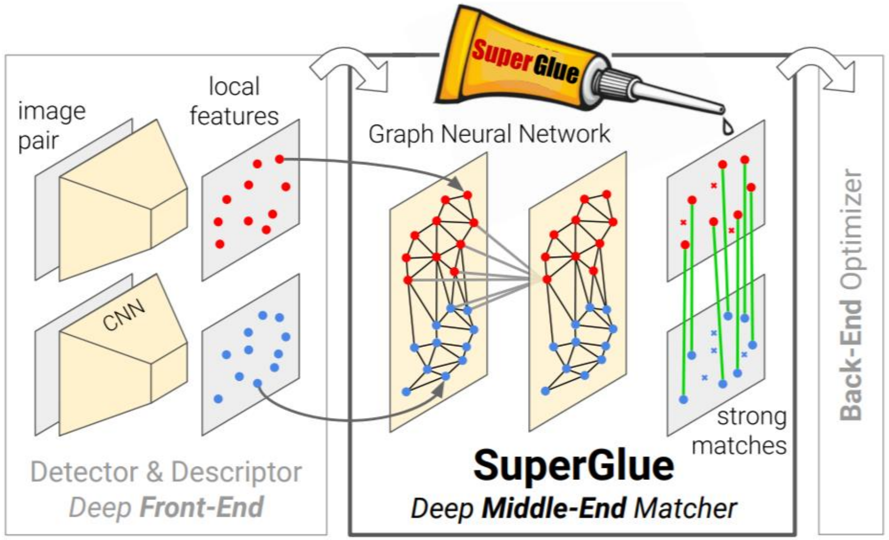
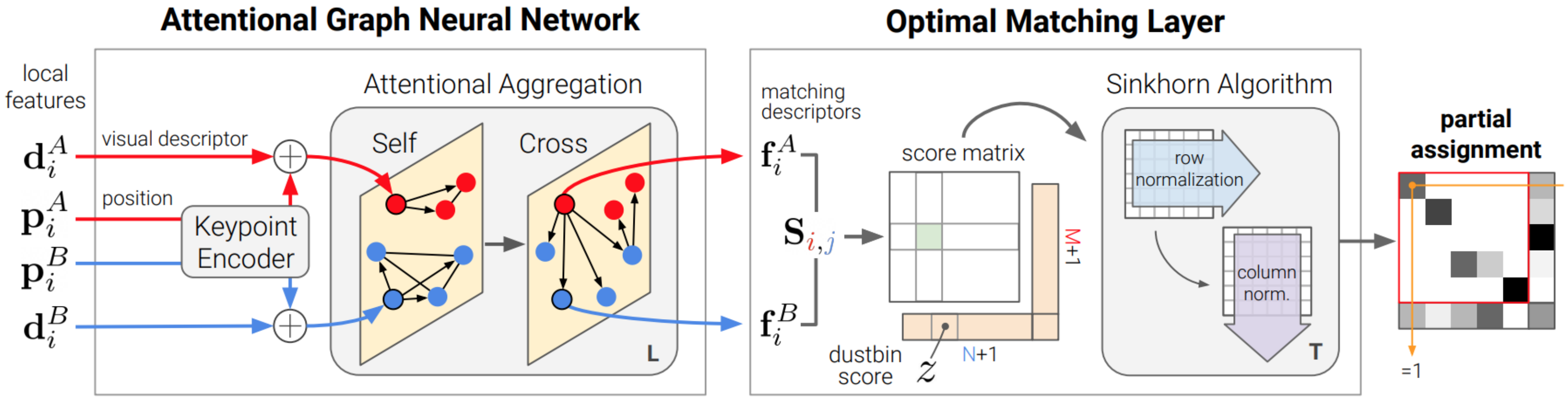
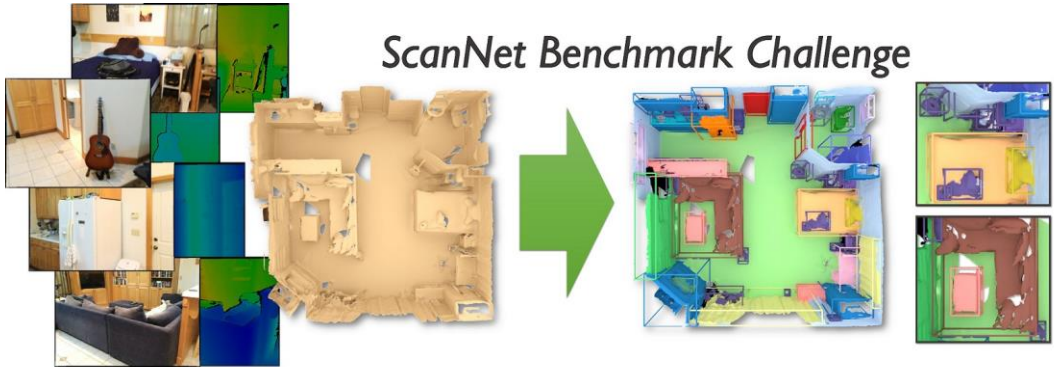


image pair
local features
CNN
Detector & Descriptor
Deep Front-End
SuperGlue
Graph Neural Network
strong matches
SuperGlue
Deep Middle-End Matcher
Back-End Optimizer

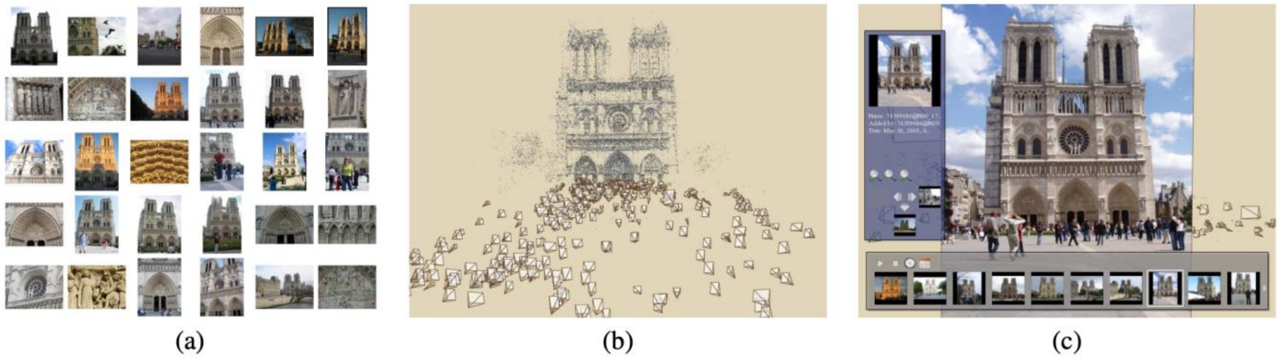

Attentional Graph Neural Network
local features
visual descriptor
position
$\mathbf{d}_i^A$
$\mathbf{p}_i^A$
$\mathbf{p}_i^B$
$\mathbf{d}_i^B$
Keypoint Encoder
Attentional Aggregation
Self
Cross
L
Optimal Matching Layer
matching descriptors
$\mathbf{f}_i^A$
$\mathbf{S}_{i,j}$
$\mathbf{f}_i^B$
score matrix
M+1
N+1
dustbin score
z
Sinkhorn Algorithm
row normalization
column norm.
T
partial assignment
=1

Experiments

- Homography Estimation
    - Sampling random homographies and applying random photometric distortions.
- Indoor Pose Estimation
    - ScanNet dataset, a large-scale indoor dataset
- Outdoor Pose Estimation
    - PhotoTourism dataset, which is part of the CVPR’19 Image Matching Challenge.



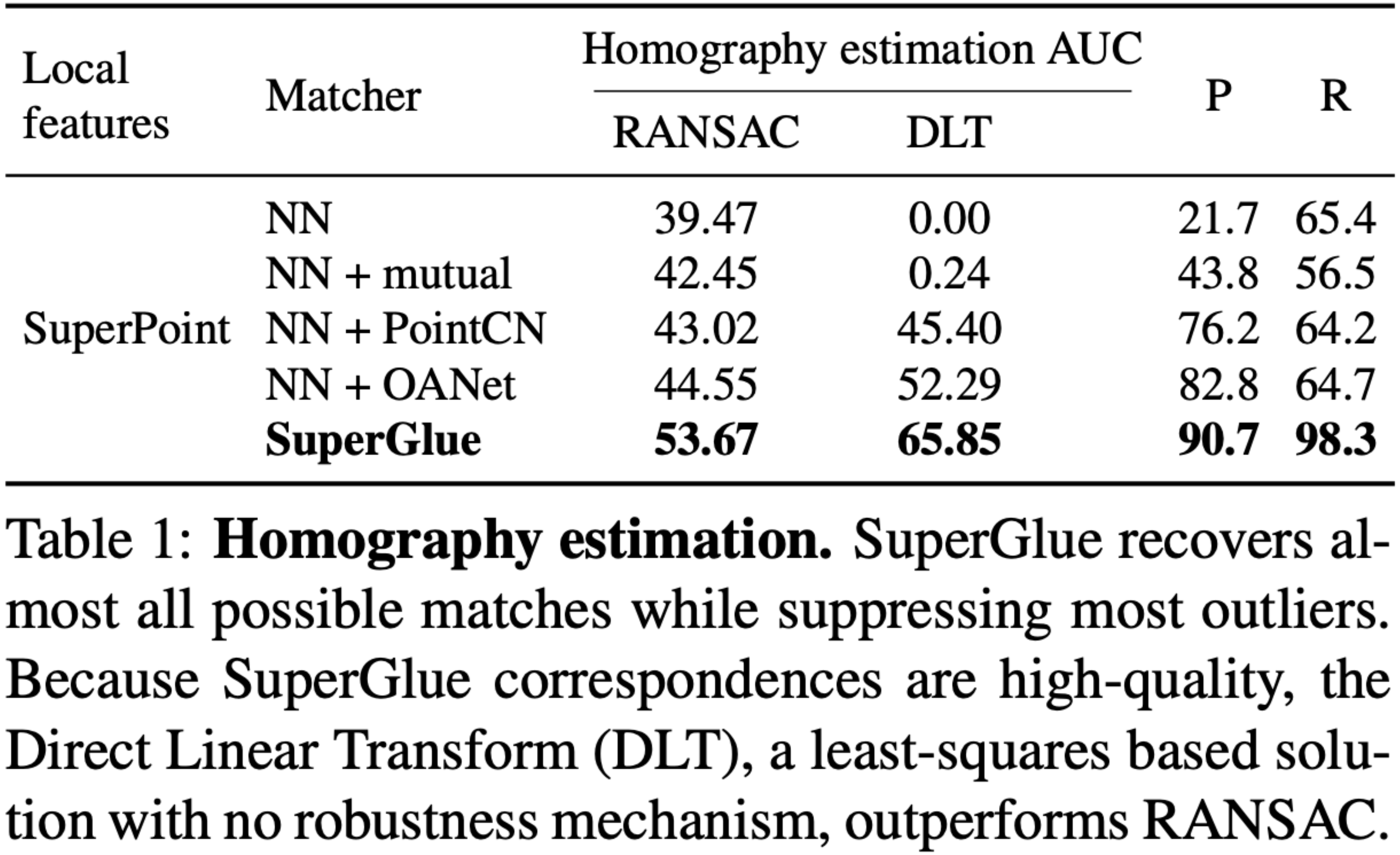
ScanNet Dataset



PhotoTourism Dataset

Dai, Angela, et al. "Scannet: Richly-annotated 3d reconstructions of indoor scenes." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
Snavely, Noah, Steven M. Seitz, and Richard Szeliski. "Photo tourism: exploring photo collections in 3D." *ACM siggraph 2006 papers*. 2006. 835-846.

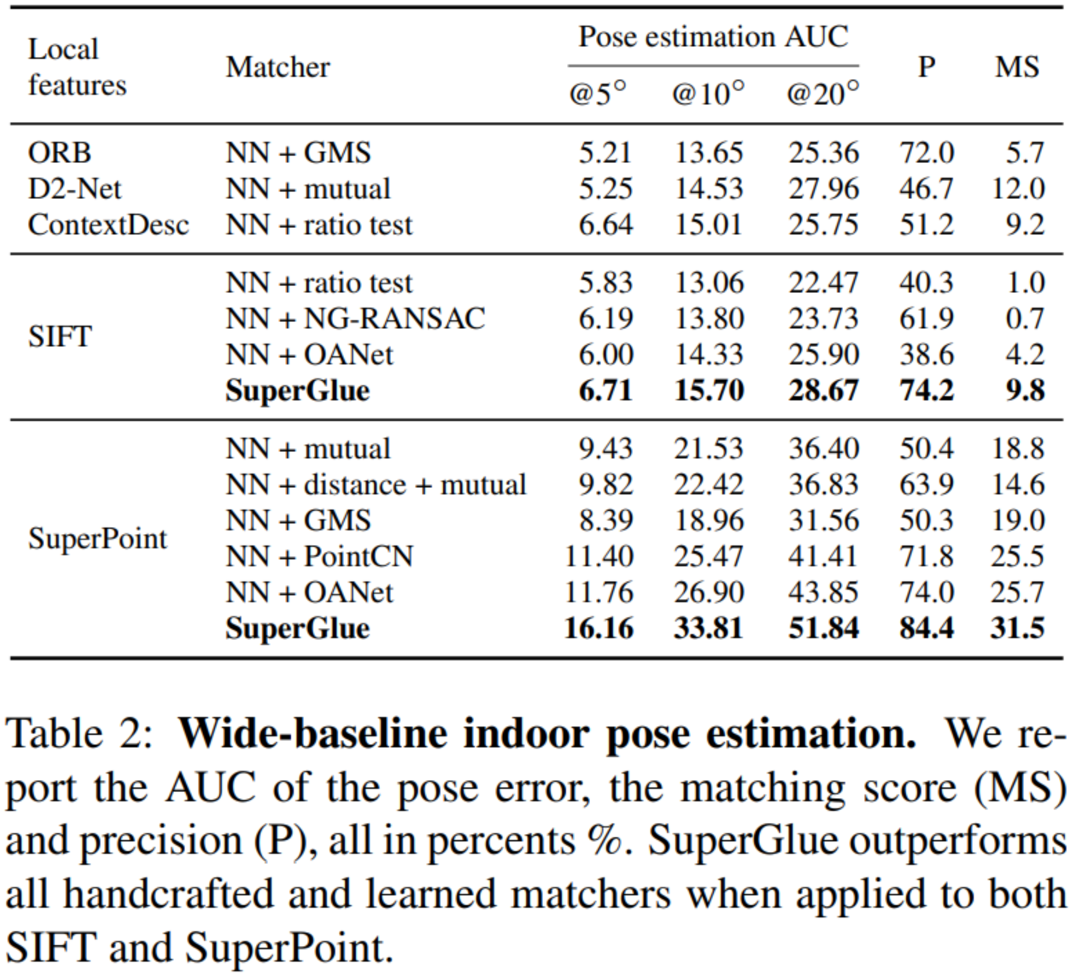
Homography Estimation



Local features	Matcher	Homography estimation AUC		P	R
		RANSAC	DLT		
SuperPoint	NN	39.47	0.00	21.7	65.4
	NN + mutual	42.45	0.24	43.8	56.5
	NN + PointCN	43.02	45.40	76.2	64.2
	NN + OANet	44.55	52.29	82.8	64.7
	SuperGlue	**53.67**	**65.85**	**90.7**	**98.3**

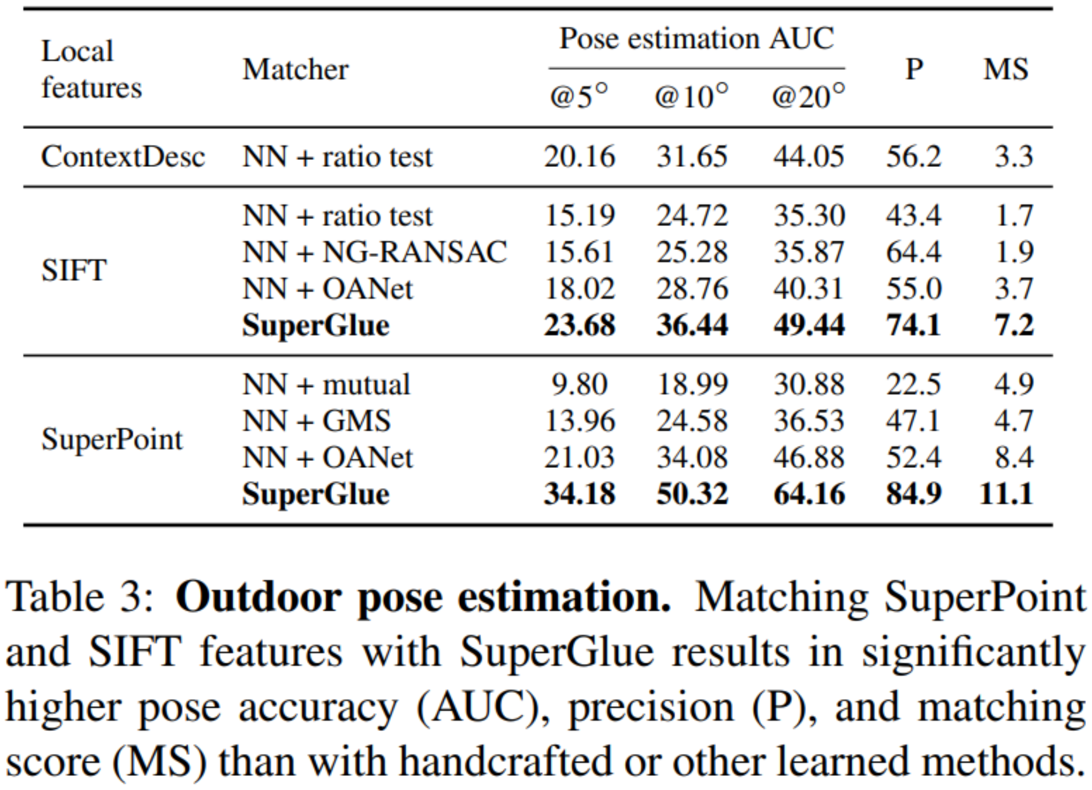
Table 1: **Homography estimation.** SuperGlue recovers almost all possible matches while suppressing most outliers. Because SuperGlue correspondences are high-quality, the Direct Linear Transform (DLT), a least-squares based solution with no robustness mechanism, outperforms RANSAC.

Indoor and Outdoor Pose Estimation



Local features	Matcher	Pose estimation AUC			P	MS
		@5°	@10°	@20°		
ORB	NN + GMS	5.21	13.65	25.36	72.0	5.7
D2-Net	NN + mutual	5.25	14.53	27.96	46.7	12.0
ContextDesc	NN + ratio test	6.64	15.01	25.75	51.2	9.2
SIFT	NN + ratio test	5.83	13.06	22.47	40.3	1.0
	NN + NG-RANSAC	6.19	13.80	23.73	61.9	0.7
	NN + OANet	6.00	14.33	25.90	38.6	4.2
	SuperGlue	**6.71**	**15.70**	**28.67**	**74.2**	**9.8**
SuperPoint	NN + mutual	9.43	21.53	36.40	50.4	18.8
	NN + distance + mutual	9.82	22.42	36.83	63.9	14.6
	NN + GMS	8.39	18.96	31.56	50.3	19.0
	NN + PointCN	11.40	25.47	41.41	71.8	25.5
	NN + OANet	11.76	26.90	43.85	74.0	25.7
	SuperGlue	**16.16**	**33.81**	**51.84**	**84.4**	**31.5**

Table 2: **Wide-baseline indoor pose estimation.** We report the AUC of the pose error, the matching score (MS) and precision (P), all in percents %. SuperGlue outperforms all handcrafted and learned matchers when applied to both SIFT and SuperPoint.



Local features	Matcher	Pose estimation AUC			P	MS
		@5°	@10°	@20°		
ContextDesc	NN + ratio test	20.16	31.65	44.05	56.2	3.3
SIFT	NN + ratio test	15.19	24.72	35.30	43.4	1.7
	NN + NG-RANSAC	15.61	25.28	35.87	64.4	1.9
	NN + OANet	18.02	28.76	40.31	55.0	3.7
	SuperGlue	**23.68**	**36.44**	**49.44**	**74.1**	**7.2**
SuperPoint	NN + mutual	9.80	18.99	30.88	22.5	4.9
	NN + GMS	13.96	24.58	36.53	47.1	4.7
	NN + OANet	21.03	34.08	46.88	52.4	8.4
	SuperGlue	**34.18**	**50.32**	**64.16**	**84.9**	**11.1**

Table 3: **Outdoor pose estimation.** Matching SuperPoint and SIFT features with SuperGlue results in significantly higher pose accuracy (AUC), precision (P), and matching score (MS) than with handcrafted or other learned methods.

Indoor and Outdoor Pose Estimation

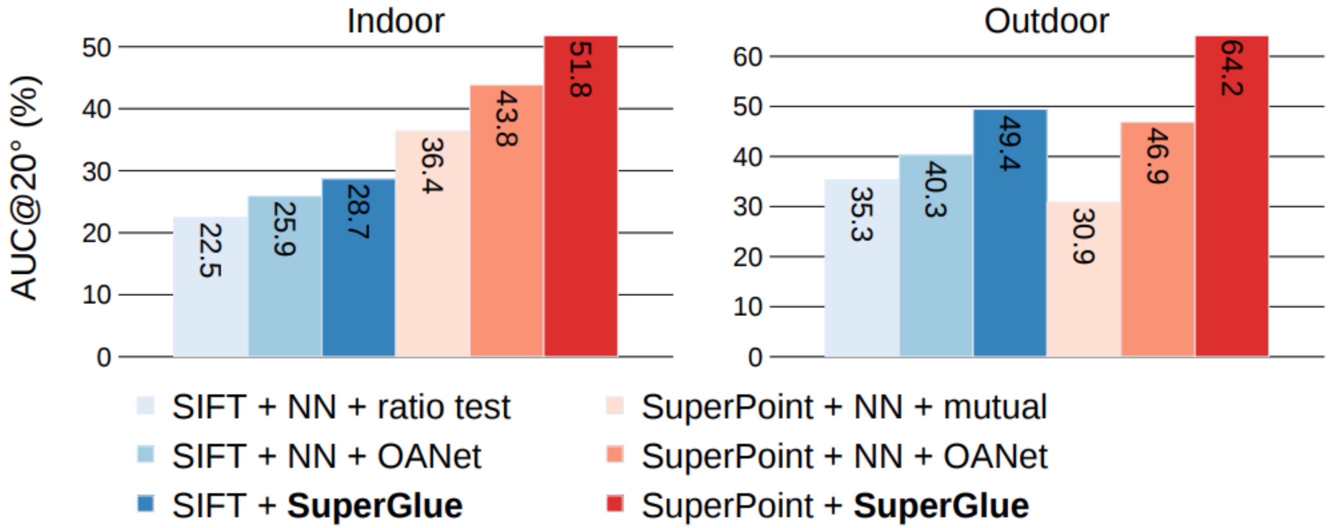
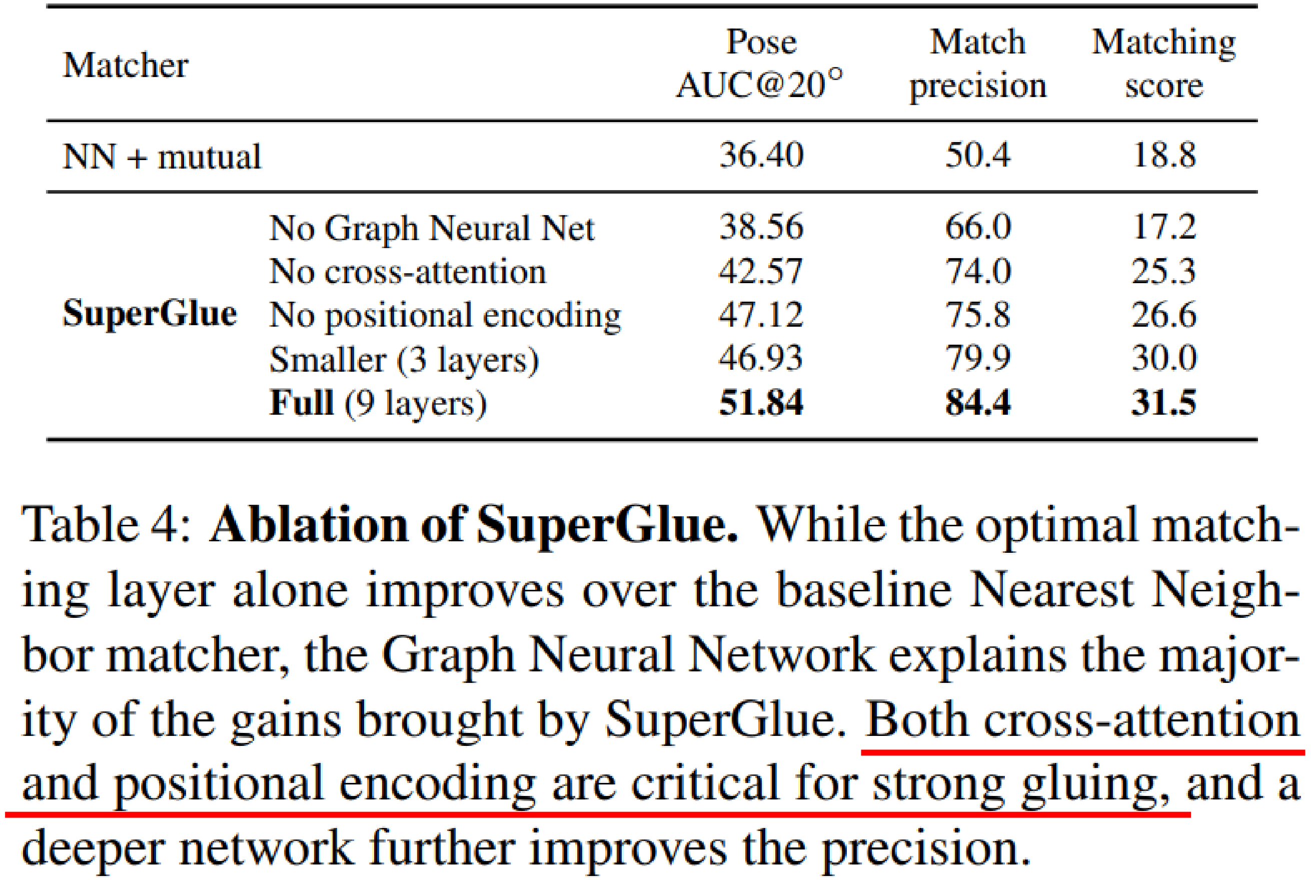


Figure 5: **Indoor and outdoor pose estimation.** SuperGlue works with SIFT or SuperPoint local features and consistently improves by a large margin the pose accuracy over OANet, a state-of-the-art outlier rejection neural network.

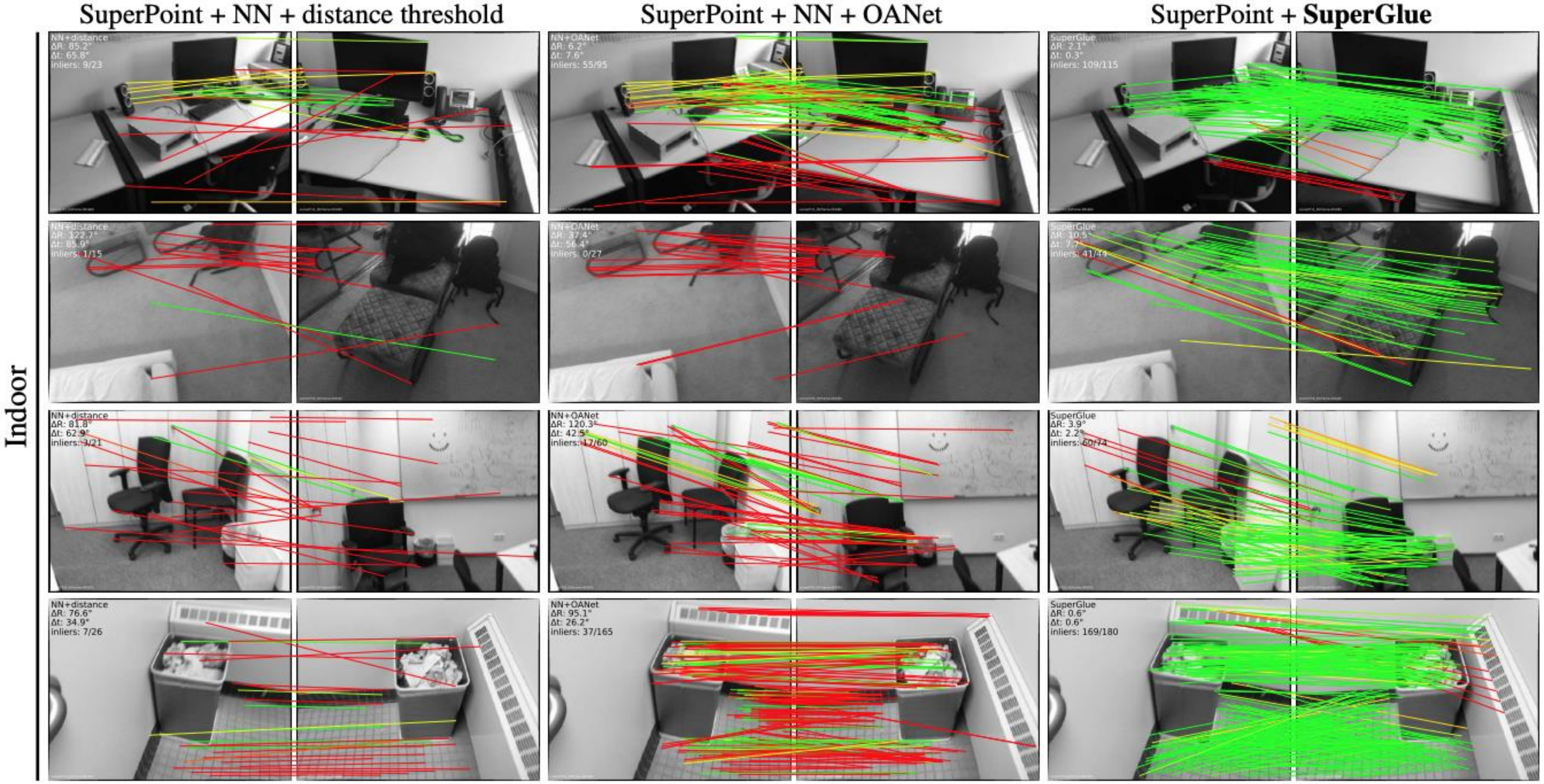
Ablation



Matcher		Pose AUC@20°	Match precision	Matching score
NN + mutual		36.40	50.4	18.8
SuperGlue	No Graph Neural Net	38.56	66.0	17.2
	No cross-attention	42.57	74.0	25.3
	No positional encoding	47.12	75.8	26.6
	Smaller (3 layers)	46.93	79.9	30.0
	Full (9 layers)	**51.84**	**84.4**	**31.5**

Table 4: **Ablation of SuperGlue.** While the optimal matching layer alone improves over the baseline Nearest Neighbor matcher, the Graph Neural Network explains the majority of the gains brought by SuperGlue. Both cross-attention and positional encoding are critical for strong gluing, and a deeper network further improves the precision.

Qualitative Results



Qualitative Results

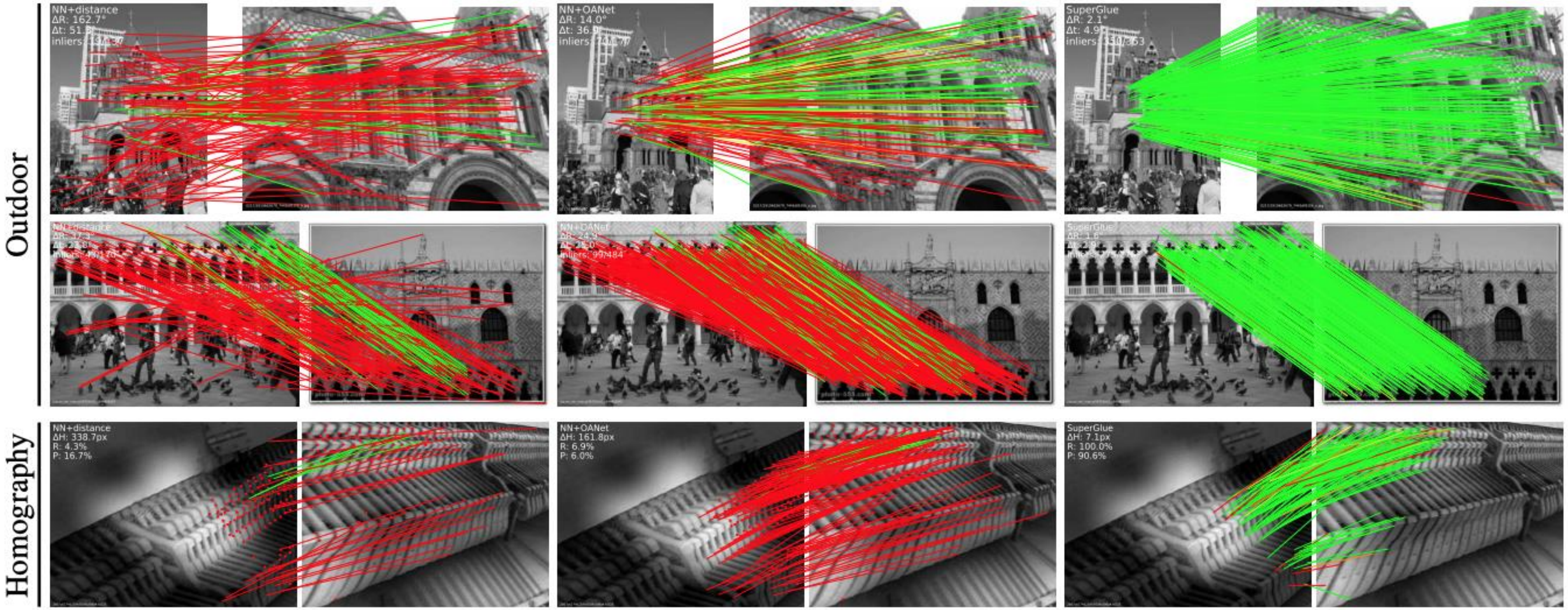


Figure 6: **Qualitative image matches.** We compare SuperGlue to the Nearest Neighbor (NN) matcher with two outlier rejectors, handcrafted and learned, in three environments. SuperGlue consistently estimates more correct matches (green lines) and fewer mismatches (red lines), successfully coping with repeated texture, large viewpoint, and illumination changes.

Day-night Pairs

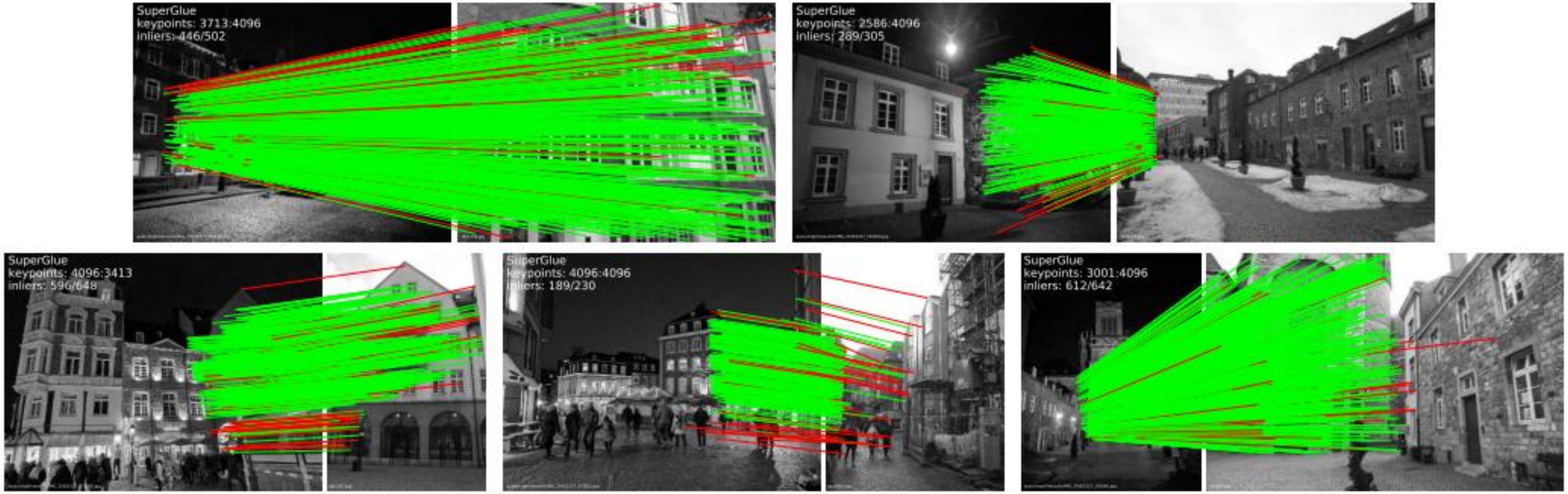


Figure 10: **Matching challenging day-night pairs with SuperGlue.** We show predicted correspondences between night-time queries and day-time databases images of the Aachen Day-Night dataset. The correspondences are colored as RANSAC inliers in green or outliers in red. Although the outdoor training 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.

Visualizing Attention

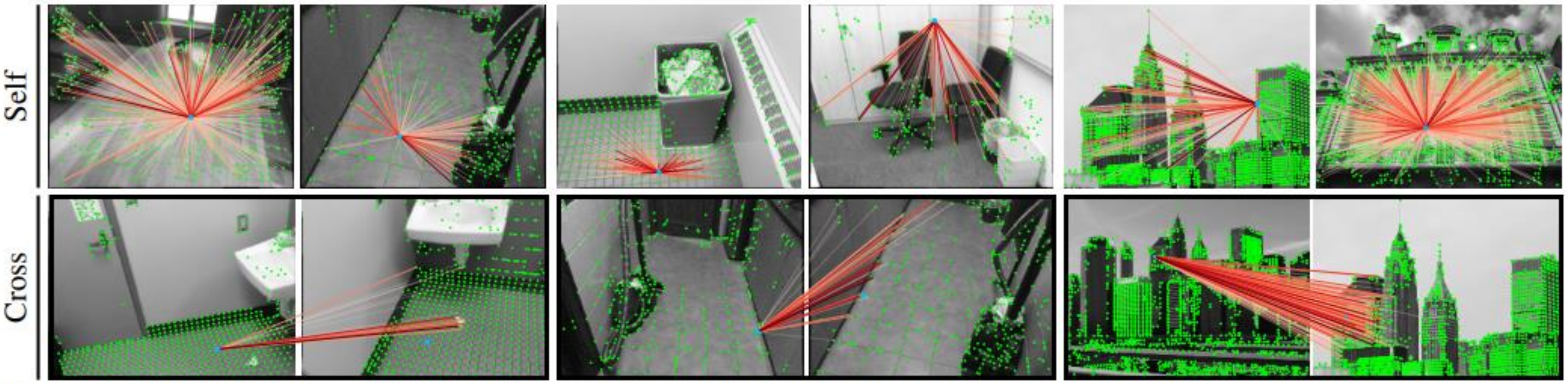


Figure 7: **Visualizing attention.** We show self- and cross-attention weights α_{ij} at various layers and heads. SuperGlue exhibits a diversity of patterns: it can focus on global or local context, self-similarities, distinctive features, or match candidates.

Failure Cases

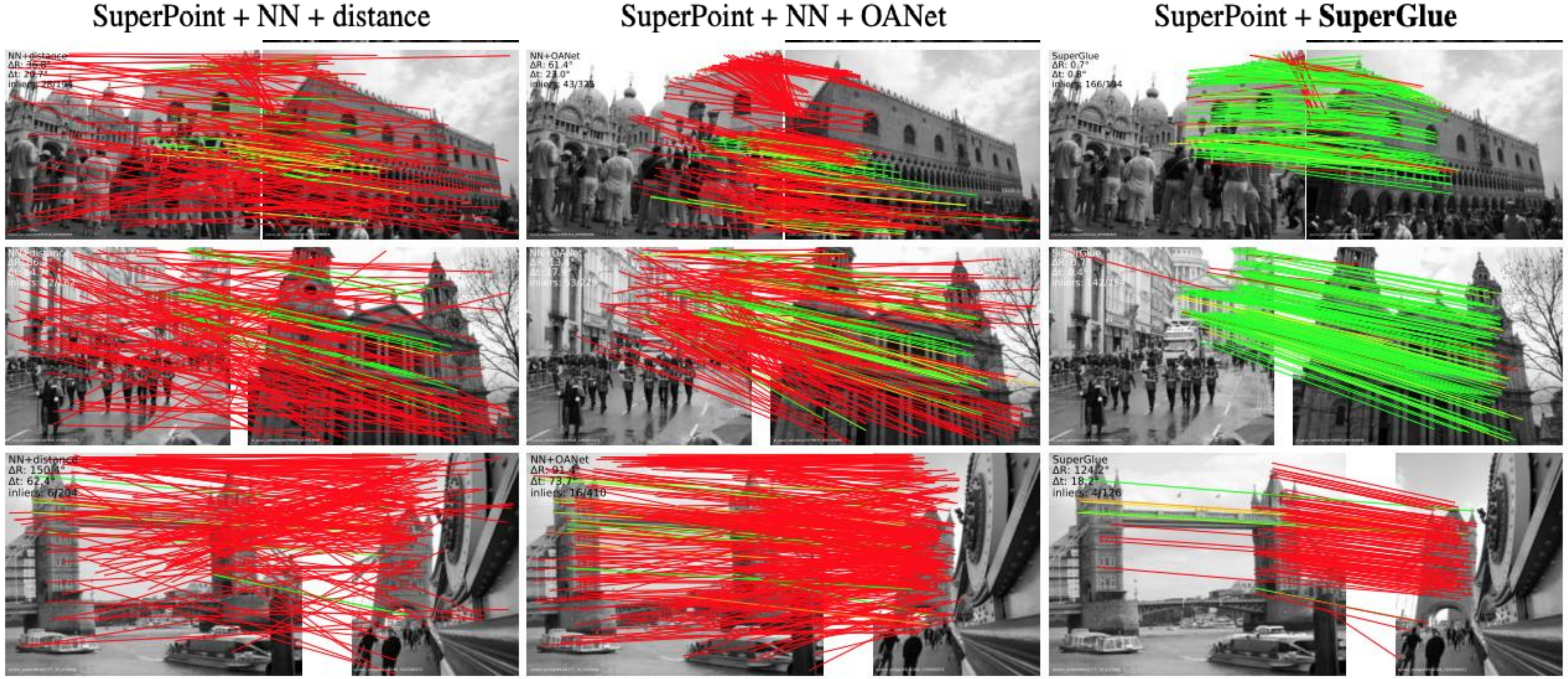
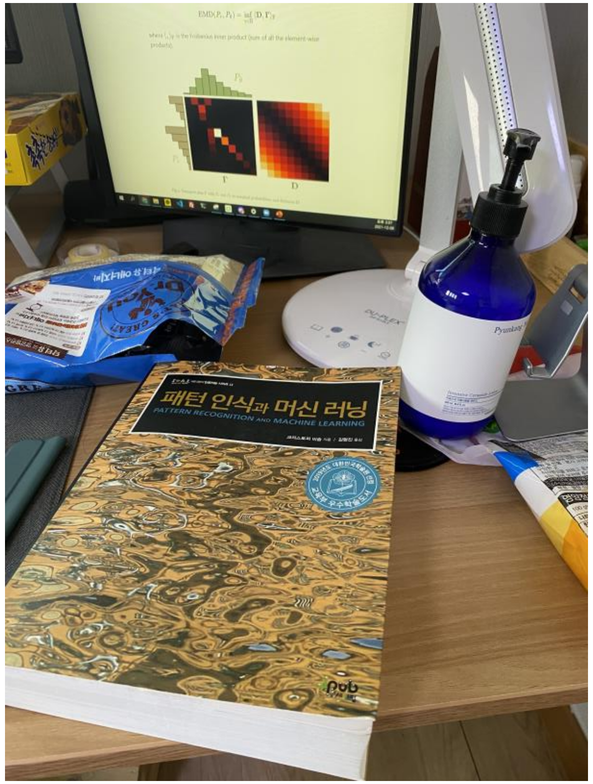
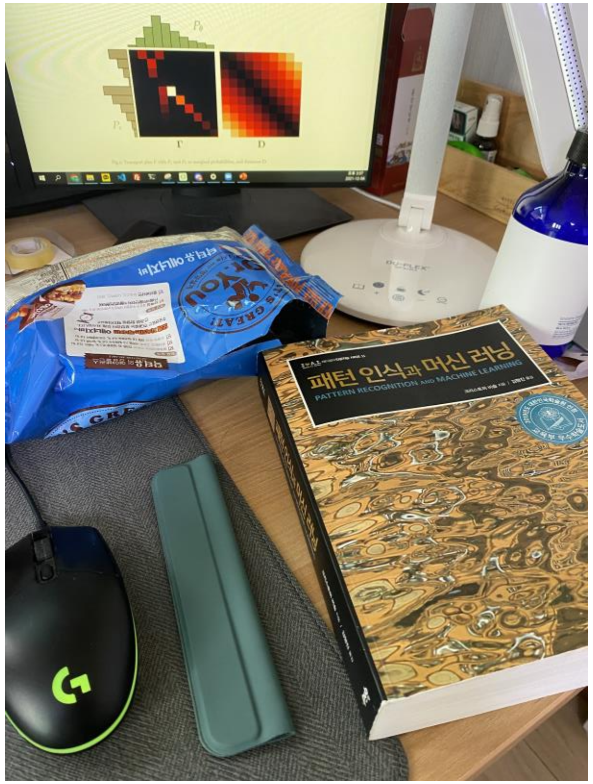


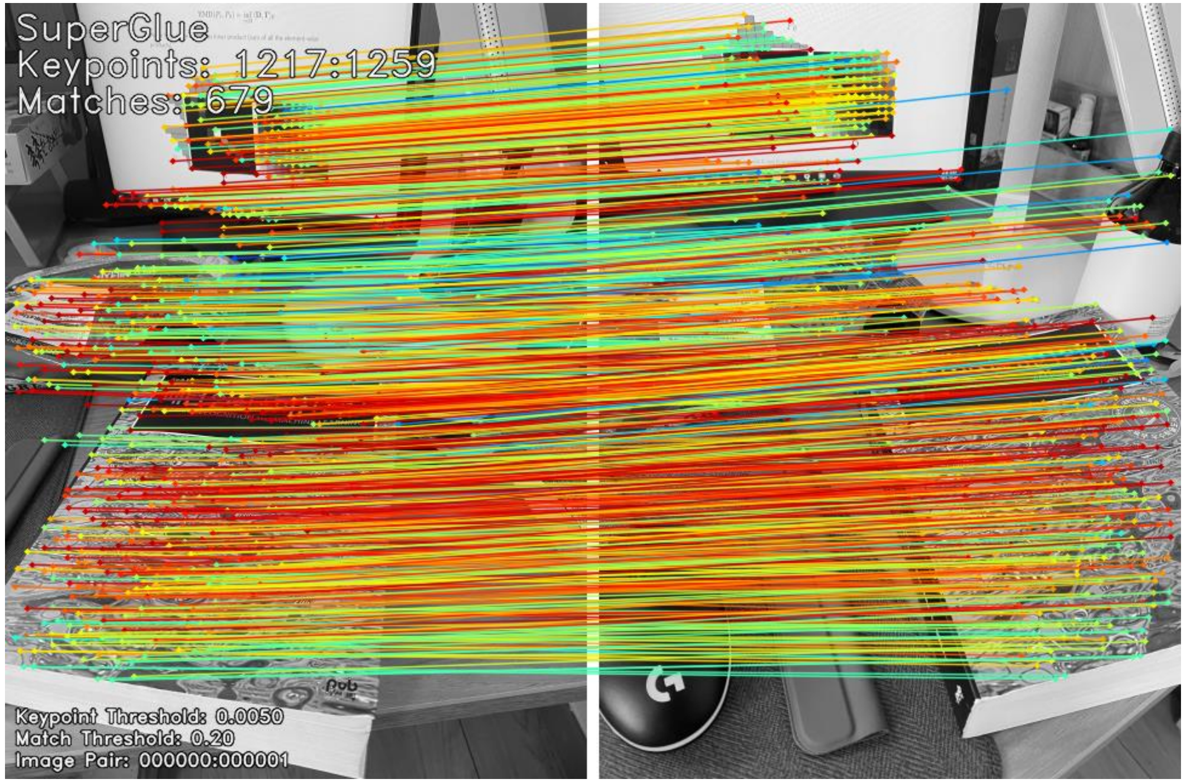
Figure 15: **More outdoor examples.** We show results on the MegaDepth validation and the PhotoTourism test sets. Correct matches are green lines and mismatches are red lines. The last row shows a failure case, where SuperGlue focuses on the incorrect self-similarity. See details in Section 5.3.

Demo

- It seems to work.







```
SuperGlue
Keypoints: 1217:1259
Matches: 679
```

```
Keypoint Threshold: 0.0050
Match Threshold: 0.20
Image Pair: 000000:000001
```

SuperGlue Github: https://github.com/magicleap/SuperGluePretrainedNetwork

Reference

- Cadena, Cesar, et al. "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age." *IEEE Transactions on robotics* 32.6 (2016): 1309-1332.
- SIFT blog post (KR): https://bskyvision.com/21
- Darkprogrammer blog post (KR), Geometry #7 Epipolar geometry, https://darkpgmr.tistory.com/83
- Vincent Herrmann blog post, Wasserstein GAN, https://vincentherrmann.github.io/blog/wasserstein/
- Dai, Angela, et al. "Scannet: Richly-annotated 3d reconstructions of indoor scenes." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- Snavely, Noah, Steven M. Seitz, and Richard Szeliski. "Photo tourism: exploring photo collections in 3D." *ACM siggraph 2006 papers*. 2006. 835-846.