End-to-end Learning of Keypoint Detector and Descriptor for Pose Invariant 3D Matching

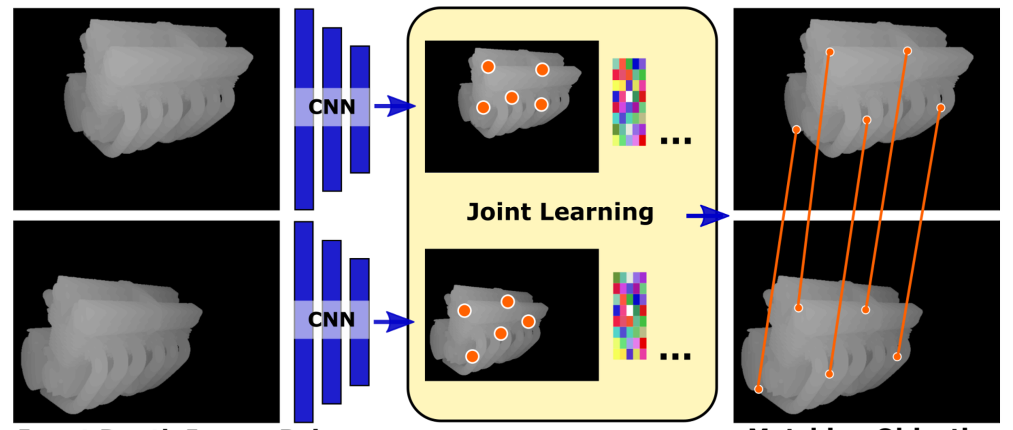
Georgios Georgakis[1], Srikrishna Karanam[2], Ziyan Wu[2], Jan Ernst[2], and Jana Kosecka[1]
George Mason University[1], Siemens Corporate Technology[2]

Problem

- Local feature learning previously considered detector and descriptor as separate objectives.
- Training requires large number of keypoint annotations.

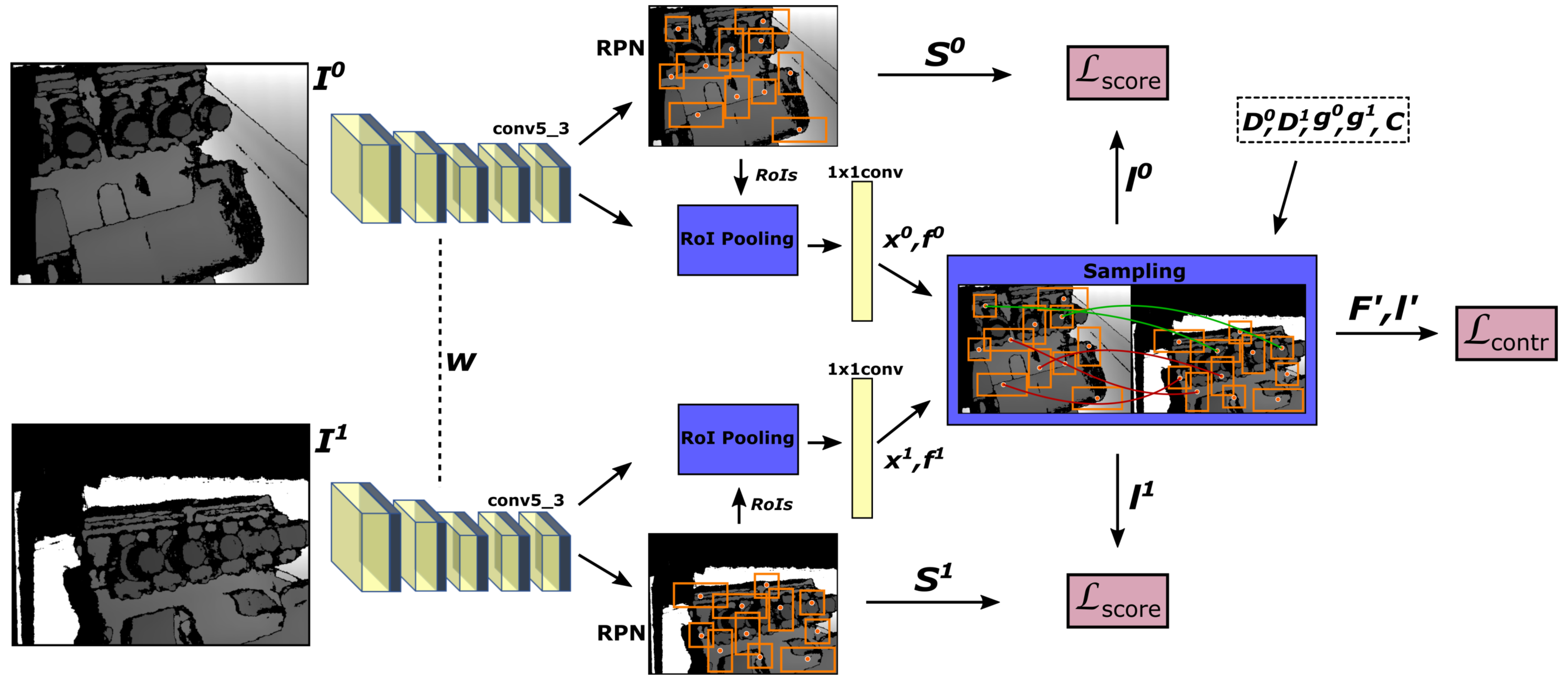
Contributions

- Joint end-to-end learning of keypoints and view-invariant patch representations.
- Generate patch correspondence on-the-fly for self-supervised training.
- Novel score loss for keypoint detection learning.



Approach

- Siamese Faster R-CNN network to bootstrap the learning process.
- Network receives as ***input*** a pair of depth images and their camera poses and ***outputs*** a set of proposals and their scores for each image.
- Sampling layer generates ground-truth pairs of patches between the two images on-the-fly based on their proximity in 3D space.
- Combination of contrastive and score losses optimize towards the matching objective.



Joint Optimization

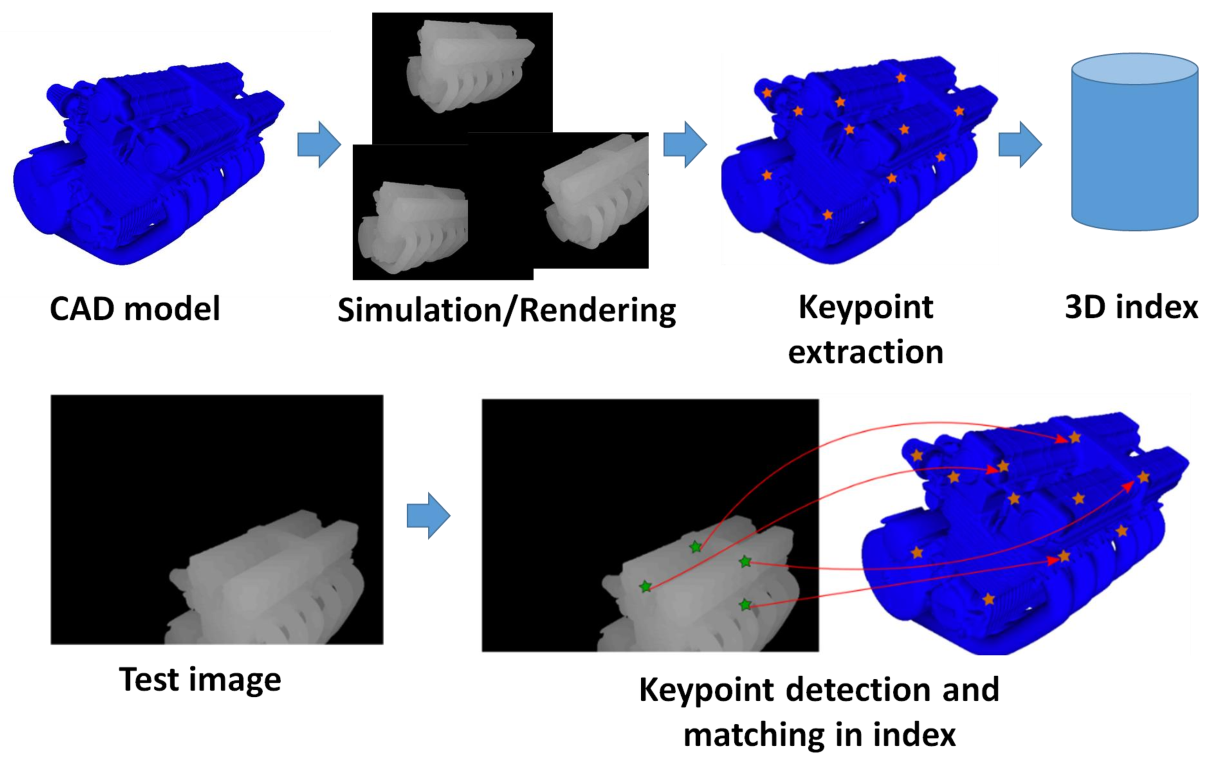
- **Contrastive loss:** Separate negative pairs and align positive pairs in feature space.

$$L_c(F', l') = \frac{\sum_{n=1}^{N} l'_n ||f_n^0 - f_n^1||^2}{2N_{pos}} + \frac{\sum_{n=1}^{N}(1 - l'_n) max(0, v - ||f_n^0 - f_n^1||)^2}{2N_{neg}}$$

- **Score loss:** Train a detector to maximize the number of correspondences between two images.

$$L_s^m(s^m, l^m) = \frac{1}{1 + N_{pos}} - \frac{\gamma \sum_{i=1}^{N} l_i^m \log s_i^m}{1 + N_{pos}}$$

Testing



- True matches decided with a small 3D distance threshold.
- **Keypoint matching accuracy:** Ratio of true matches to all matches.

References

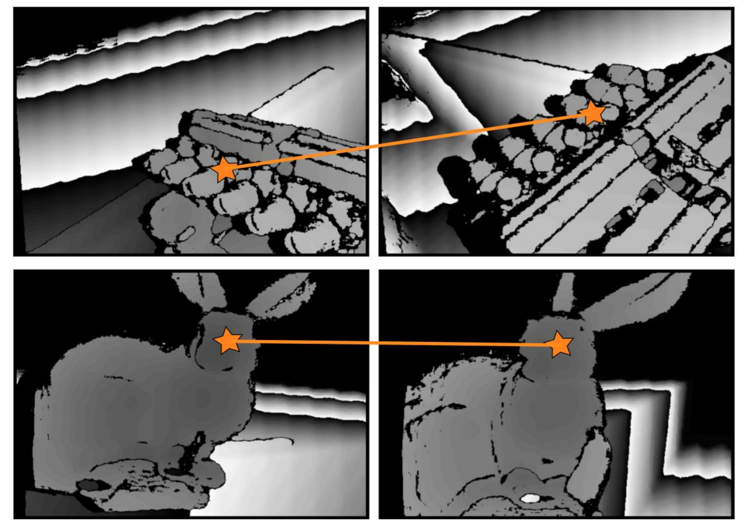
1. A. Zeng et al, 3dmatch: Learning local geometric descriptors from RGB-D reconstructions, CVPR, 2017.
2. S. Salti et al, Learning a descriptor –specific 3D keypoint detector, ICCV, 2015.
3. K. M. Yi et al, LIFT: Learned invariant feature transform, ECCV, 2016.
4. P. Wohlhart, V. Lepetit, Learning descriptors for object recognition and 3D pose estimation, CVPR, 2015.
5. B. Planche et al, Depthsynth: Real-time realistic synthetic data generation from CAD models for 2.5D recognition, 3DV, 2017.

Experimental Results

- Noise-free and noisy 3D models created by adding simulated sensor noise using Depthsynth[5].
- Comparison to hand-crafted descriptors and descriptor learning baselines.
- Model aclearns to generate keypoints in non-noisy areas.
- Learned representation demonstrates viewpoint-invariance.
- All tables show keypoint matching curacy.

Noisy Stanford 3D models

Method	Armadillo	Bunny	Dragon	Buddha	Average
ISS+SHOT	0.8	0.5	0.6	0.4	0.6
ISS+FPFH	2.0	1.7	2.4	1.4	1.9
Harris3D+SHOT	8.0	11.4	6.9	6.7	8.3
KPL+SHOT	18.0	12.8	15.4	9.1	13.8
Harris3D+FPFH	14.5	16.0	16.4	10.5	14.4
Harris3D+3DMatch	14.9	17.7	27.8	15.1	18.8
Ours-No-Score	10.0	18.3	25.2	12.5	16.5
Ours	**25.2**	**31.9**	**45.7**	**27.7**	**32.6**

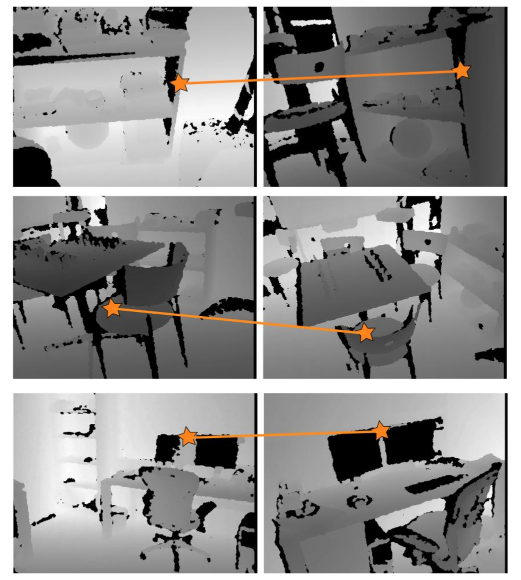


Engine 3D model

Method	Noise-Free	Noisy
ISS+SHOT	47.9	0.5
KPL+SHOT	57.2	2.8
ISS+FPFH	61.1	2.9
Harris3D+SHOT	60.1	5.9
Harris3D+FPFH	**79.1**	12.8
Harris3D+3DMatch	66.2	20.7
Ours-Rnd	29.8	7.3
Ours-No-Score	40.7	11.1
Ours-Transfer	-	17.8
Ours	67.4	**23.8**

MSR-7 Scenes

Method	Accuracy
ISS+SHOT	23.0
ISS+FPFH	24.3
Harris3D+FPFH	37.4
Harris3D+SHOT	37.9
Harris3D+3DMatch	38.2
Ours	**41.2**



Keypoint Detection