Learning Deep Convolutional Frontends for Visual SLAM

Tomasz Malisiewicz, PhD
Research @ Magic Leap

September 14th, 2018: ECCV 2018 Geometry Meets Deep Learning Workshop
September 24th, 2018: Data Science Warsaw Meetup

MagicLeap One Headshots from @YouTube

Focus of today's talk

- Applied Computer Vision Research @ Magic Leap
- Intersection of Deep Learning and SLAM

#1

SuperPoint: Self-Supervised Interest Point Detection and Description

Daniel DeTone
Magic Leap
Sunnyvale, CA
ddetone@magicleap.com

Tomasz Malisiewicz
Magic Leap
Sunnyvale, CA
tmalisiewicz@magicleap.com

Andrew Rabinovich
Magic Leap
Sunnyvale, CA
arabinovich@magicleap.com

Abstract

This paper presents a self-supervised framework for training interest point detectors and descriptors suitable for a large number of multiple-view geometry problems in computer vision. As opposed to patch-based neural networks, our fully-convolutional model operates on full-sized images and jointly computes pixel-level interest point locations and associated descriptors in one forward pass. We introduce Homographic Adaptation, a multi-scale, multi-homography approach for boosting interest point detection repeatability and performing cross-domain adaptation (e.g., synthetic-to-real). Our model, when trained on the MS-COCO generic image dataset using Homographic Adaptation, is able to repeatedly detect a much richer set of interest points than the initial pre-adapted deep model and any other traditional corner detector. The final system gives rise to state-of-the-art homography estimation results on HPatches when compared to LIFT, SIFT and ORB.

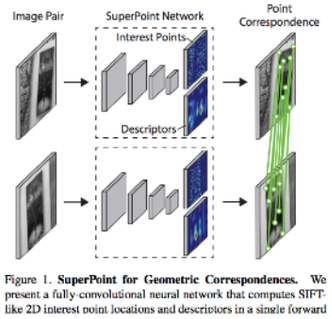


Figure 1. **SuperPoint for Geometric Correspondences.** We present a fully-convolutional neural network that computes SIFT-like 2D interest point locations and descriptors in a single forward

CVPR 2018

#2

GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks

Zhao Chen [1] **Vijay Badrinarayanan** [1] **Chen-Yu Lee** [1] **Andrew Rabinovich** [1]

Abstract

Deep multitask networks, in which one neural network produces multiple predictive outputs, can offer better speed and performance than their single-task counterparts but are challenging to train properly. We present a gradient normalization (GradNorm) algorithm that automatically balances training in deep multitask models by dynamically tuning gradient magnitudes. We show that such as smartphones, wearable devices, and robots/drones. Such a system can be enabled by multitask learning, where one model shares weights across multiple tasks and makes multiple inferences in one forward pass. Such networks are not only scalable, but the shared features within these networks can induce more robust regularization and boost performance as a result. In the ideal limit, we can thus have the best of both worlds with multitask networks: more efficiency and higher performance.

ICML 2018

#3

Estimating Depth from RGB and Sparse Sensing

Zhao Chen, Vijay Badrinarayanan, Gilad Drozdov, and Andrew Rabinovich

Magic Leap, Inc.
{zchen, vbadrinarayanan, gdrozdov, arabinovich}@magicleap.com

Abstract. We present a deep model that can accurately produce dense depth maps given an RGB image with known depth at a very sparse set of pixels. The model works *simultaneously* for both indoor/outdoor scenes and produces state-of-the-art dense depth maps at nearly real-time speeds on both the NYUv2 and KITTI datasets. We surpass the state-of-the-art for monocular depth estimation even with depth values for only 1 out of every $\sim$ 10000 image pixels, and we outperform other sparse-to-dense depth methods at all sparsity levels. With depth values for 1/256 of the image pixels, we achieve a mean error of less than 1% of actual depth on indoor scenes, comparable to the performance of consumer-grade depth sensor hardware. Our experiments demonstrate that it would indeed be possible to efficiently transform sparse depth measurements obtained using e.g. lower-power depth sensors or SLAM systems into high-quality dense depth maps.

Keywords: Sparse-to-Dense Depth, Depth Estimation, Deep Networks.

ECCV 2018

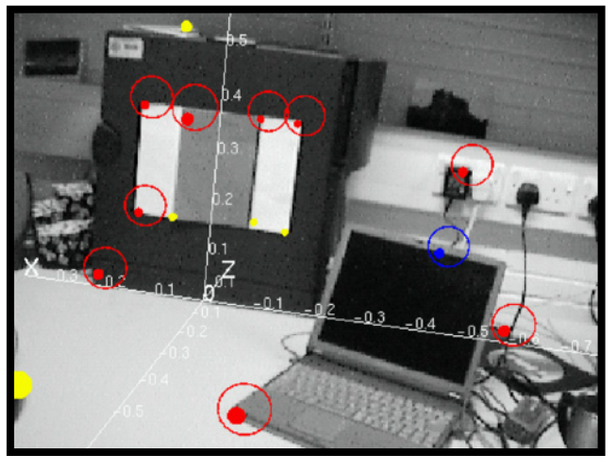
Today's Main Ideas

- "SuperPoint"
  - A Deep SLAM Frontend
  - Multi-task fully convolutional network
- "Homographic Adaptation"
  - Self-supervised recipe to train keypoints
  - Homography-inspired domain adaptation
- Snapshot of Deep Learning Research @ Magic Leap
  - GradNorm for Multi-task learning (ICML 2018)
  - Deep Depth Densification (ECCV 2018)

2000-2015 Visual SLAM

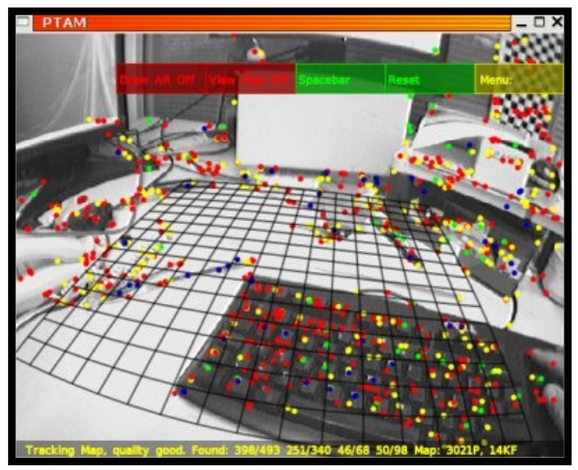
- Great Visual SLAM Research
- Real-time systems emerge
- Very few learned components

MonoSLAM



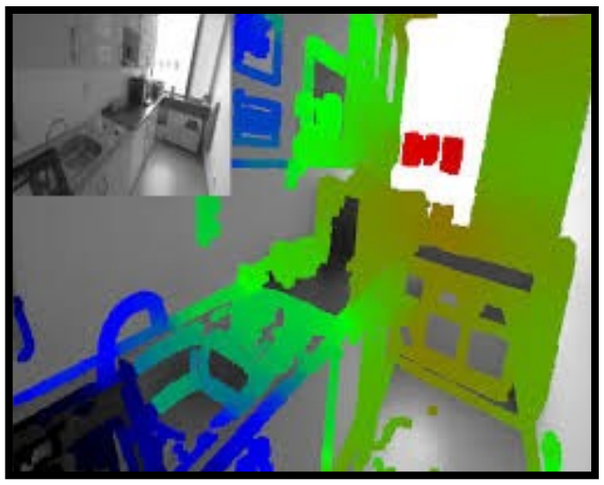
DTAM

PTAM



ElasticFusion

LSD-SLAM



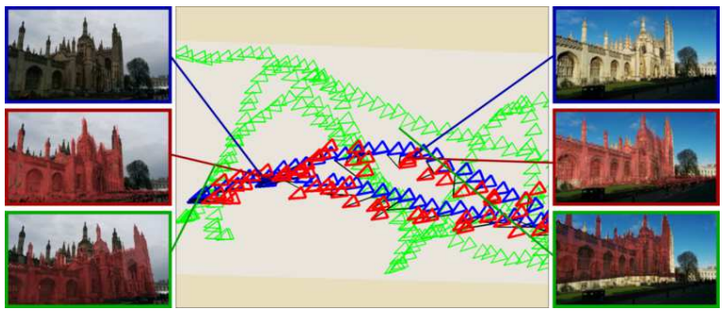
KinectFusion

Event-camera SLAM

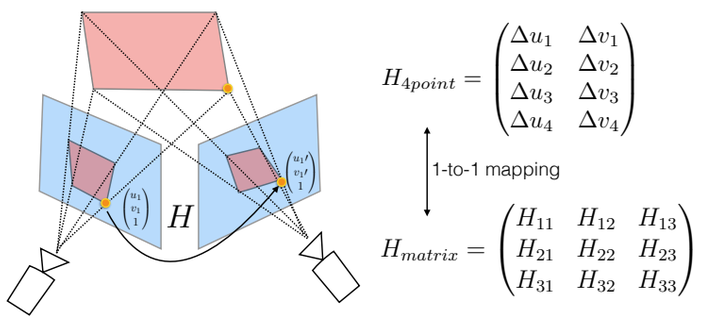
DynamicFusion

Collage courtesy: Andrew Davison's ICCV 2015 Future of Real-time SLAM workshop talk

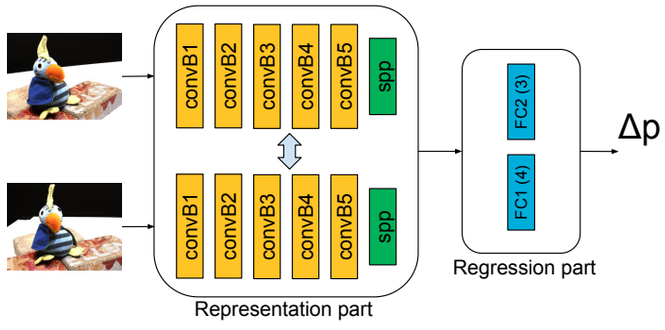
2015-2016: Simple End-to-End Deep SLAM?



PoseNet: A Convolutional Network for Real-time 6 DOF Localization



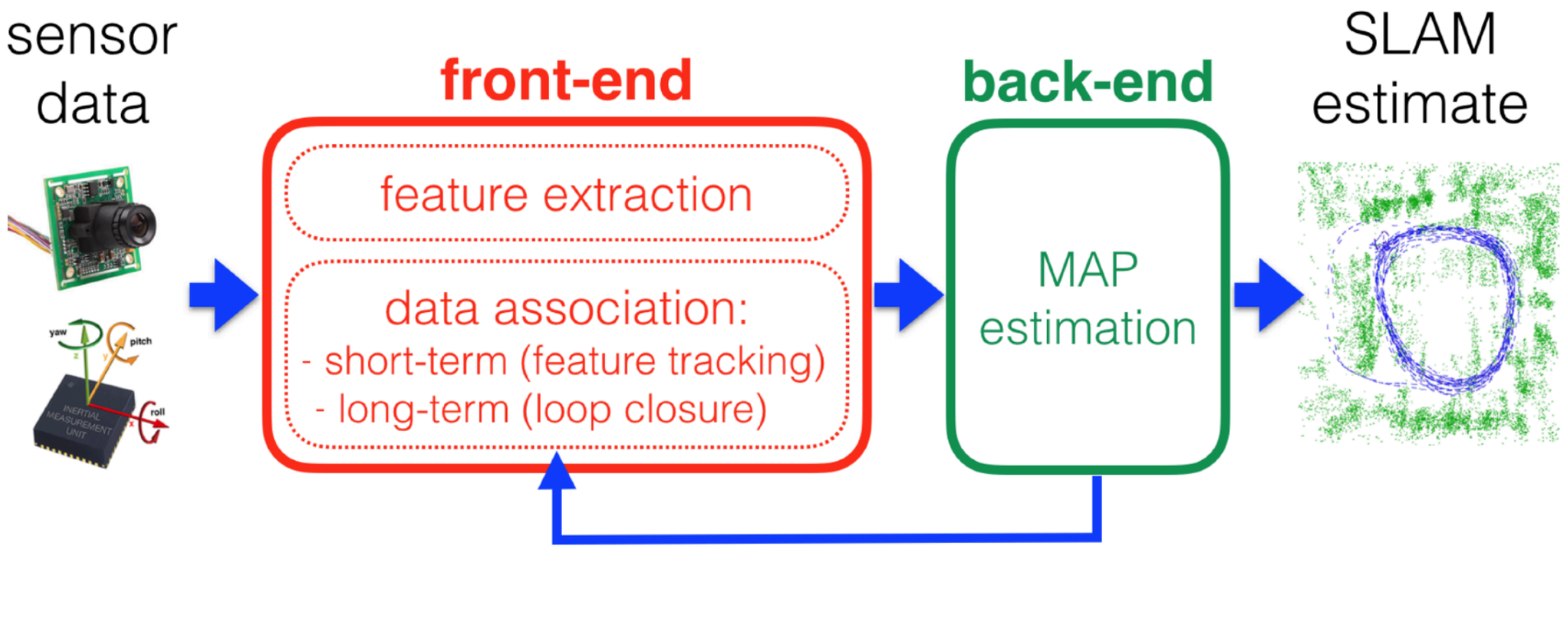
Deep Image Homography Estimation



Relative Camera Pose Estimation Using Convolutional Neural Networks

- Deep Learning excitement is very high
- Simple end-to-end setups work across many computer vision tasks
  - Purely data-driven, powerful
  - Very few heuristics / little hand-tuning
- Accuracy not yet competitive
  - Maybe due to lack of large-scale data

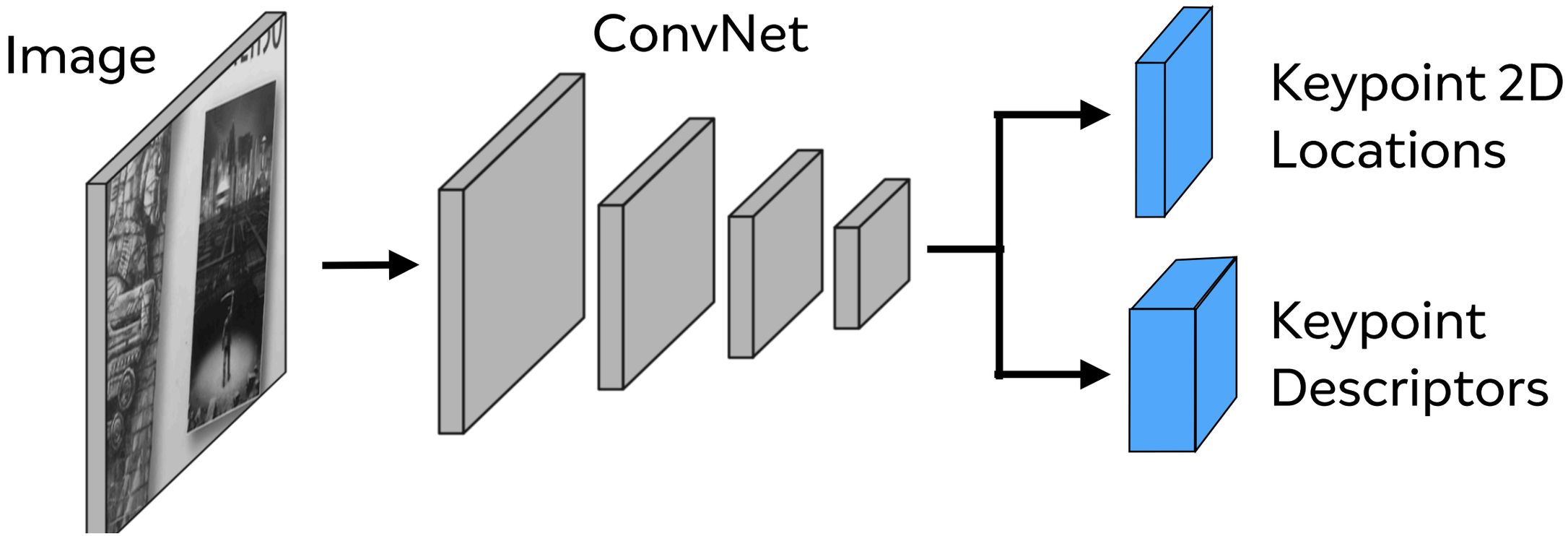
2017-2018: Splitting Up the Problem



- **Frontend**: Image inputs
  - Deep Learning success: Images + ConvNets
- **Backend**: Optimization over pose and map quantities
  - Use Bundle Adjustment

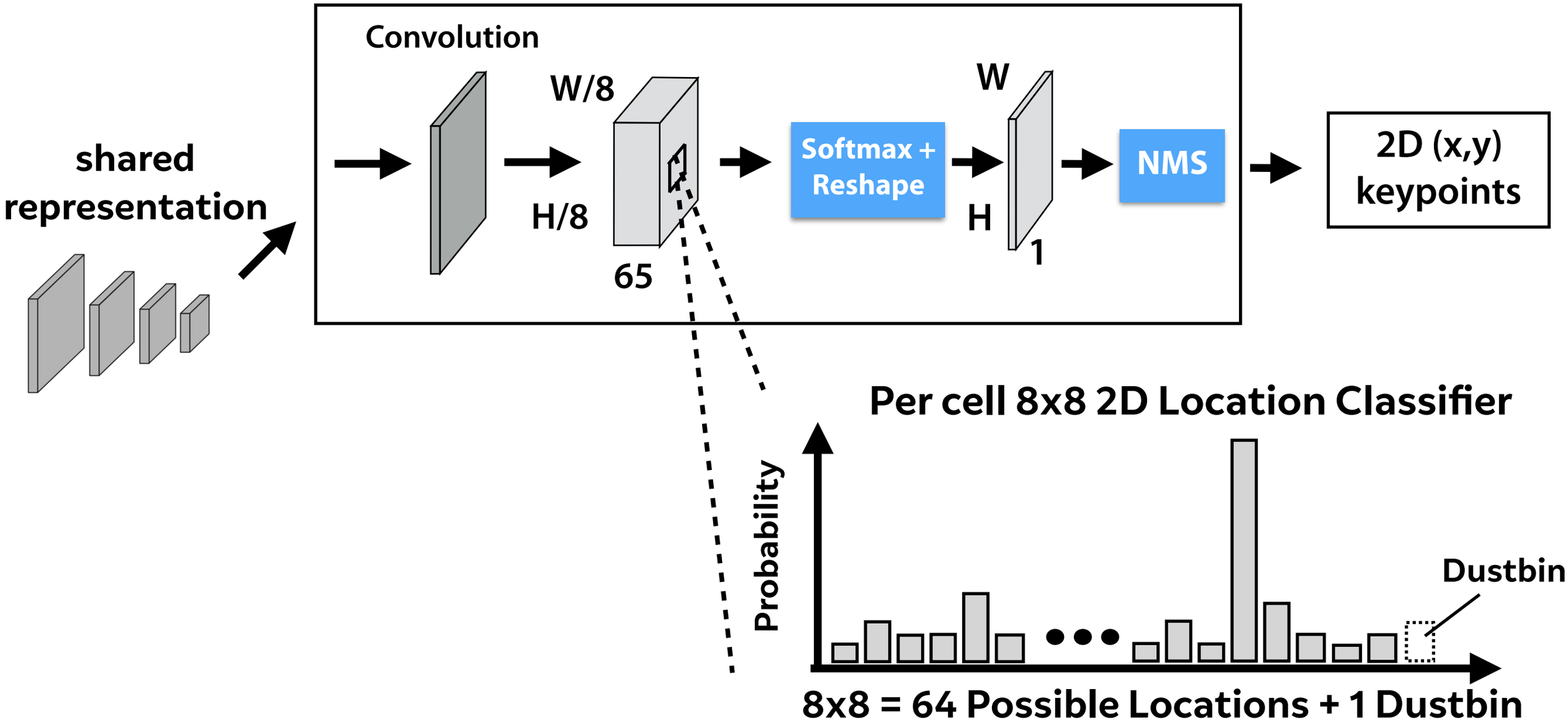
Photo Credit: Cadena et al 2016

SuperPoint: A Deep SLAM Front-end



- Powerful fully convolutional design
  - Points + descriptors computed jointly
  - Share VGG-like backbone
- Designed for real-time
  - Tasks share ~90% of compute
  - Two learning-free decoders: no deconvolution layers

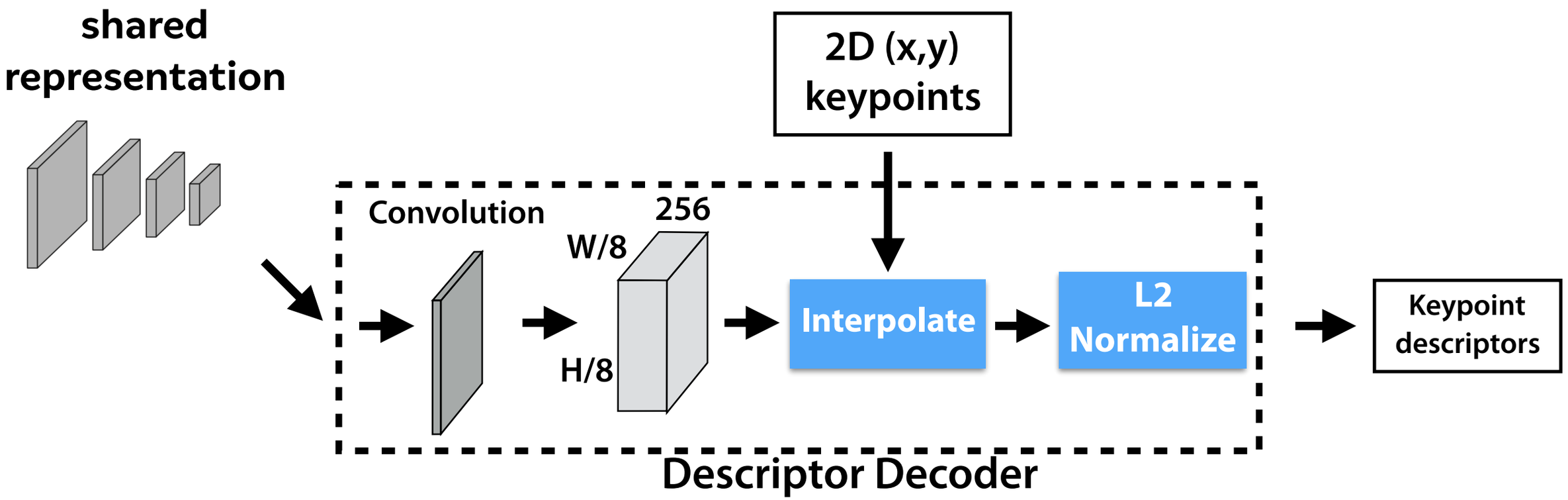
Keypoint / Interest Point Decoder



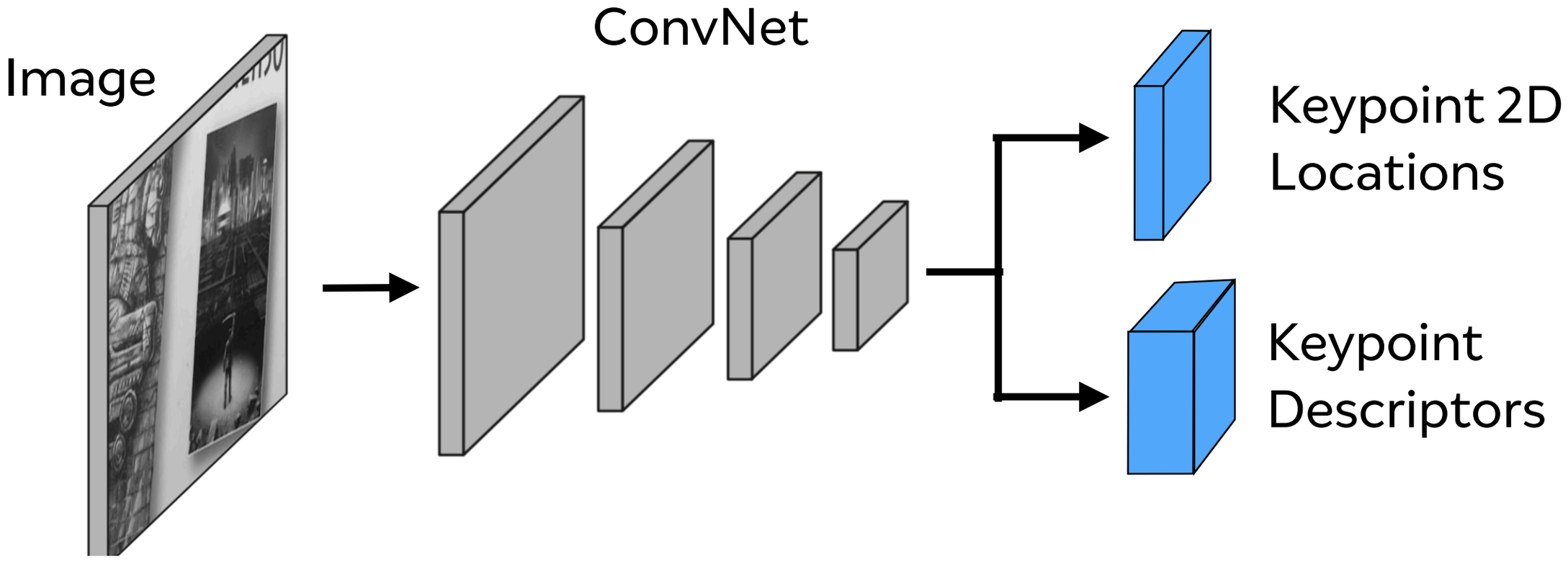
- No deconvolution layers
- Each output cell responsible for local 8x8 region

Descriptor Decoder

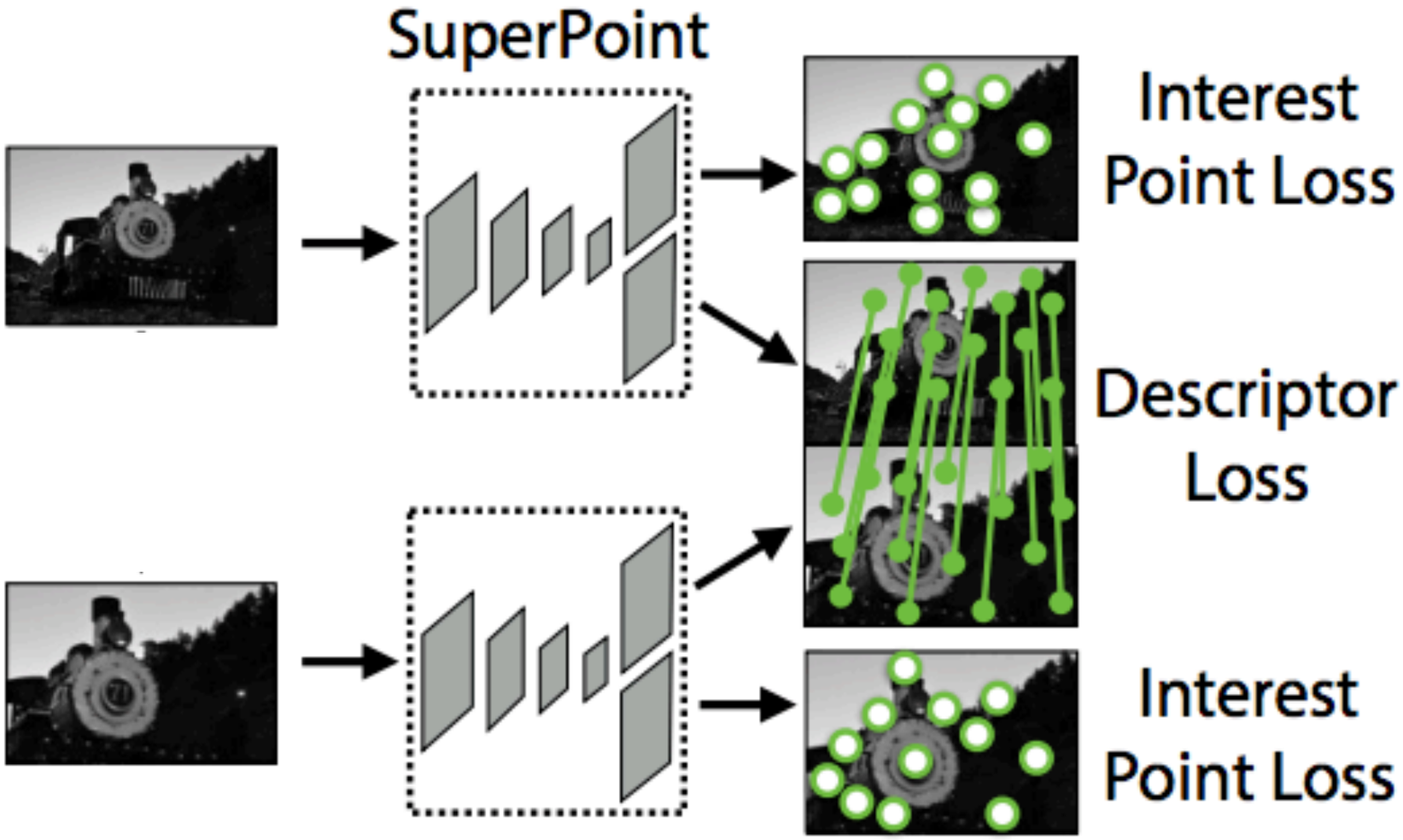
- Also no deconvolution layers
- Interpolate using 2D keypoint into coarse descriptor map



How To Train SuperPoint?



Setting up the Training



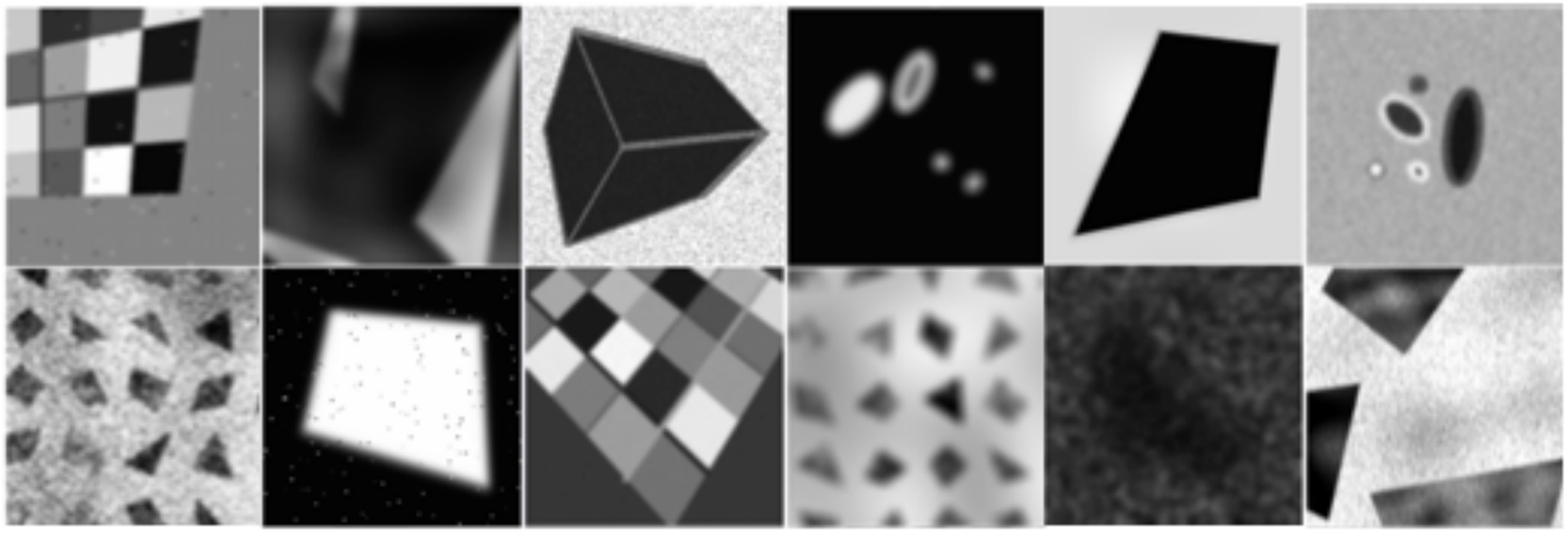
- Siamese training -> pairs of images
- Descriptor trained via metric learning
- Keypoints trained via supervised keypoint labels

How to get Keypoint Labels for Natural Images?

- Need large-scale dataset of annotated images
  - Too hard for humans to label

Self-Supervised Approach

Synthetic Shapes (has interest point labels)



First train on this

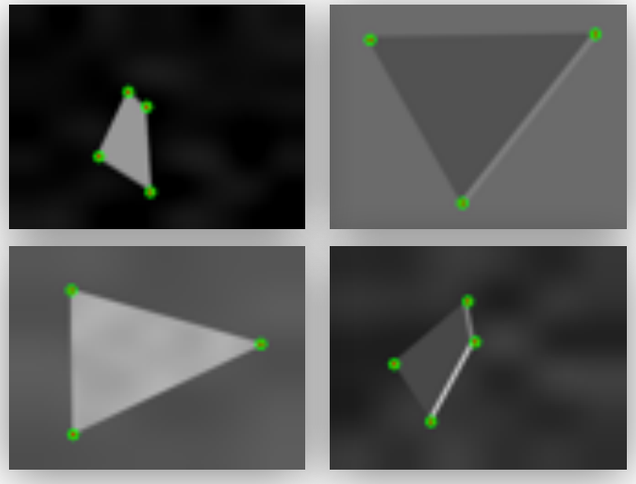
"Homographic Adaptation"

MS-COCO (no interest point labels)

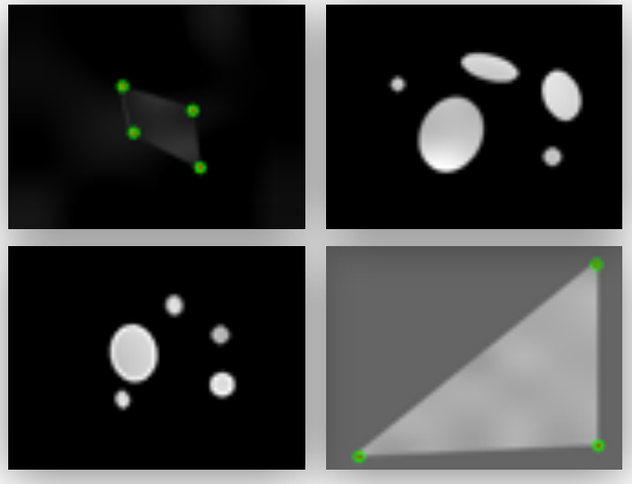
Use resulting detector to label this

Synthetic Training

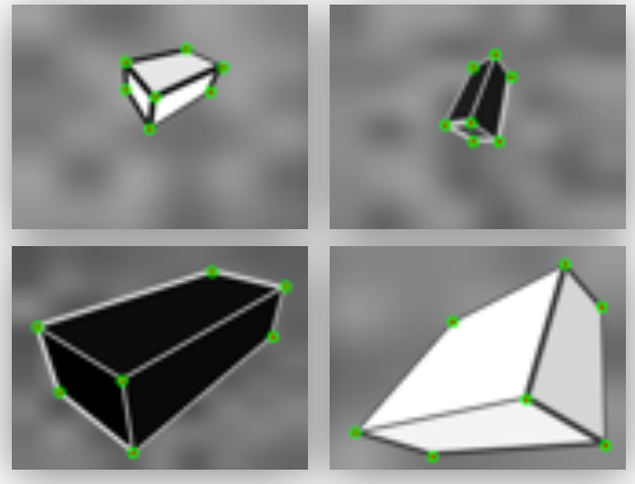
- Non-photorealistic shapes
- Heavy noise
- Effective and easy



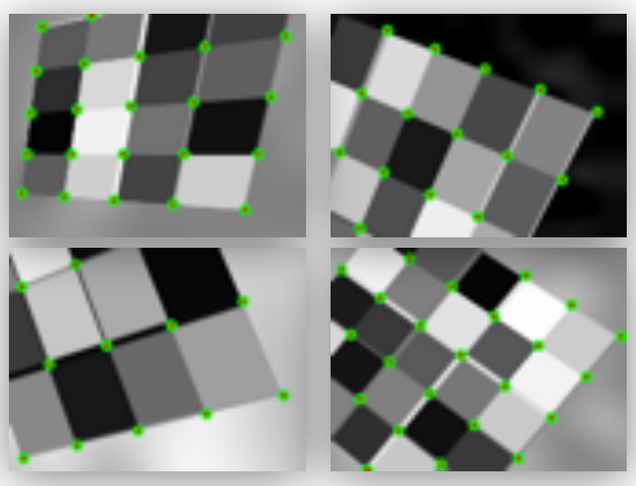
Quads/Tris



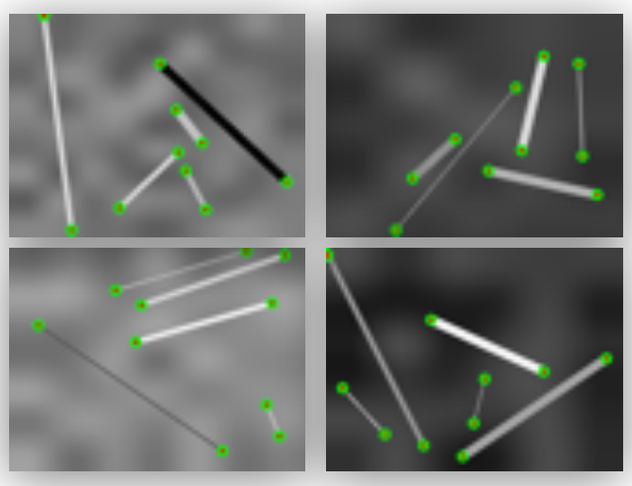
Quads/Tris/Ellipses



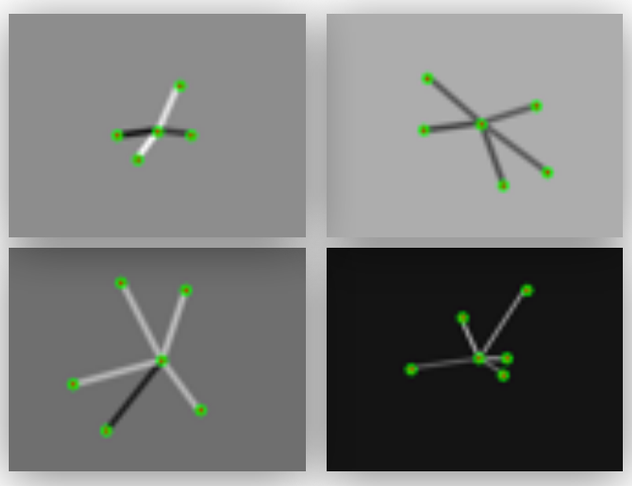
Cubes



Checkerboards

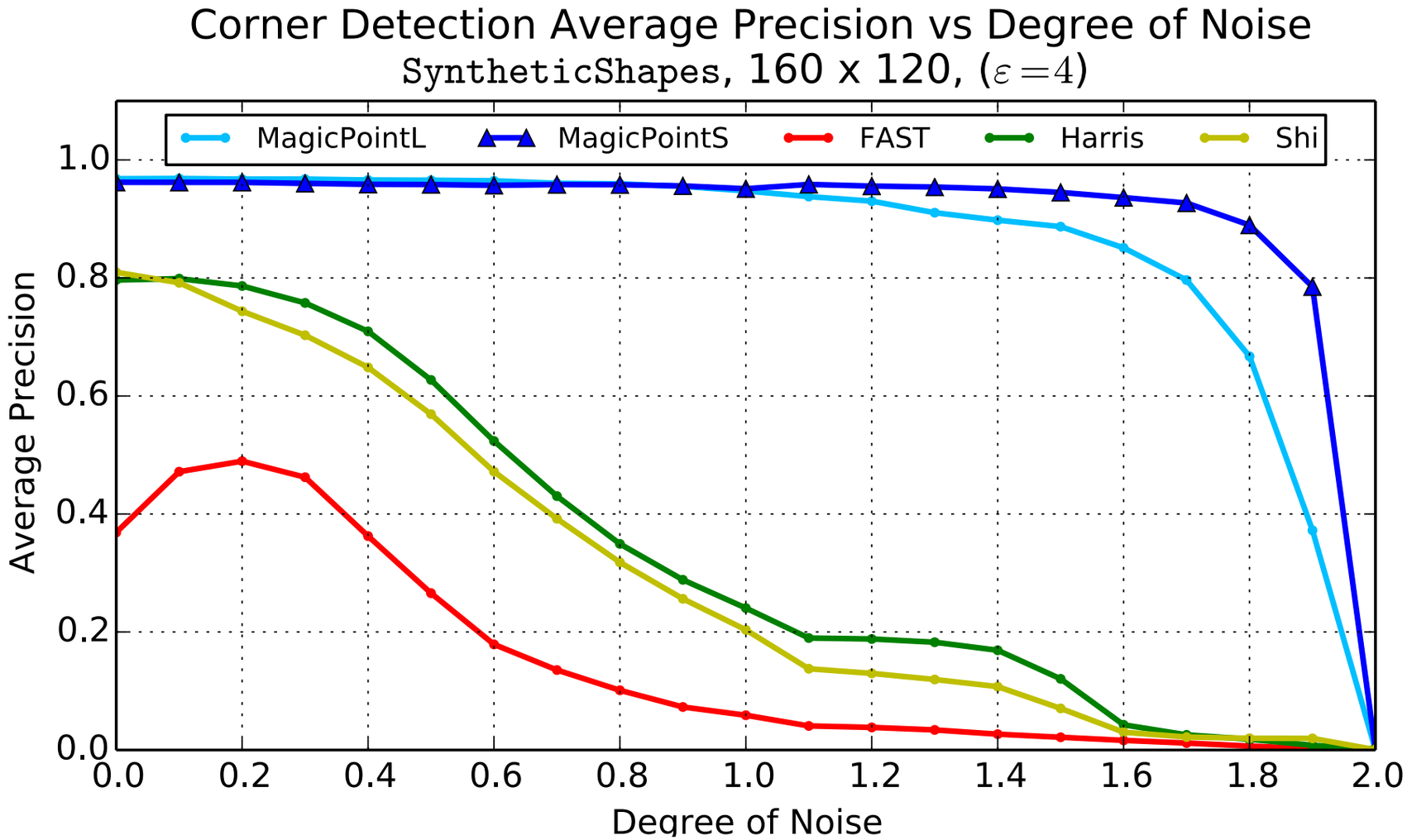


Lines

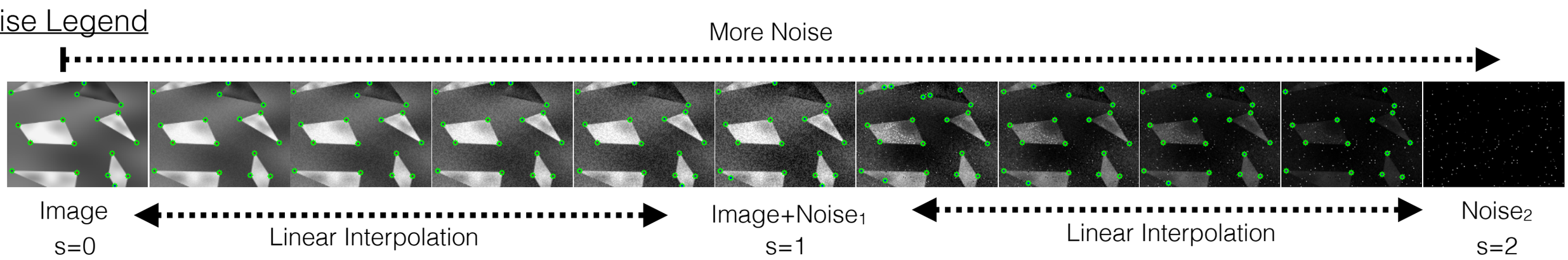


Stars

Early Version of SuperPoint (MagicPoint)



Noise Legend



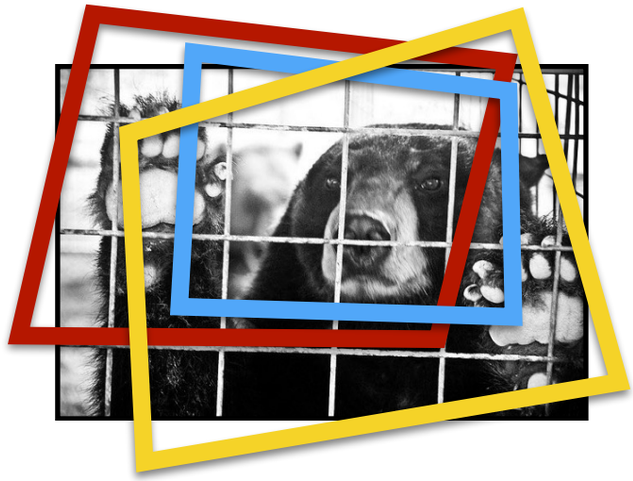
DeTone, D., Malisiewicz, T., Rabinovich, A. Toward Geometric DeepSLAM. In arXiv: 1707.07410.

Generalizing to Real Data

- Synthetically trained detector
  - Works! Despite large domain gap
    - Worked well on geometric structures
    - Under performed on certain textures unseen during training

Unlabeled Input Image

Synthetic Warp + Run Detector



Homographic Adaptation

Point Set #1

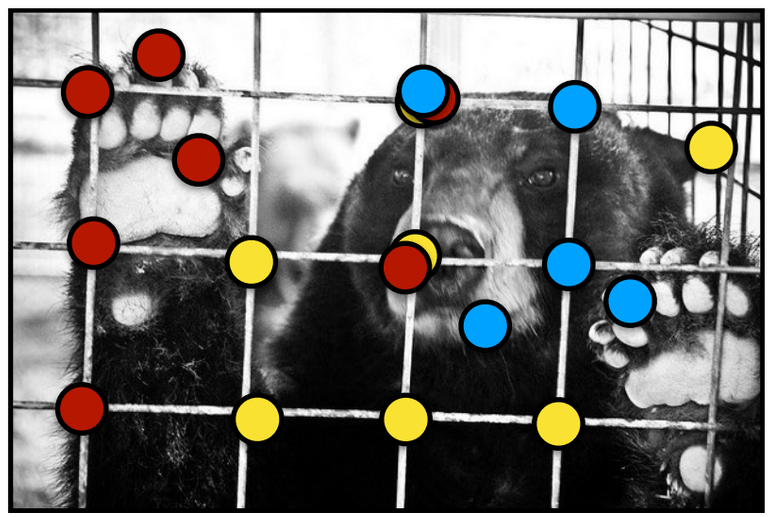
Point Set #2

Point Set #3

- Simulate planar camera motion with homographies
- Self-labelling technique
  - Suppress spurious detections
  - Enhance repeatable points

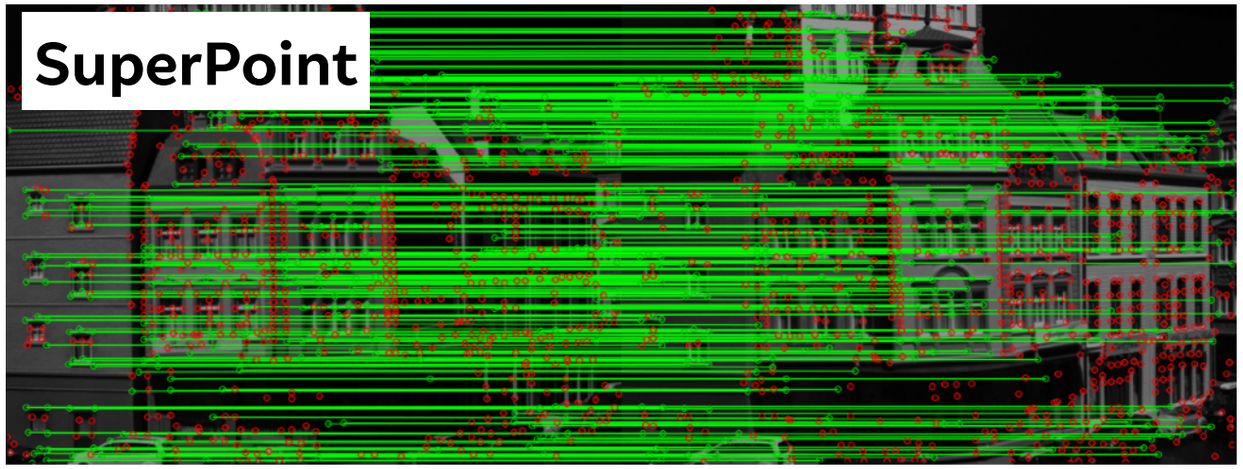
Point Aggregation

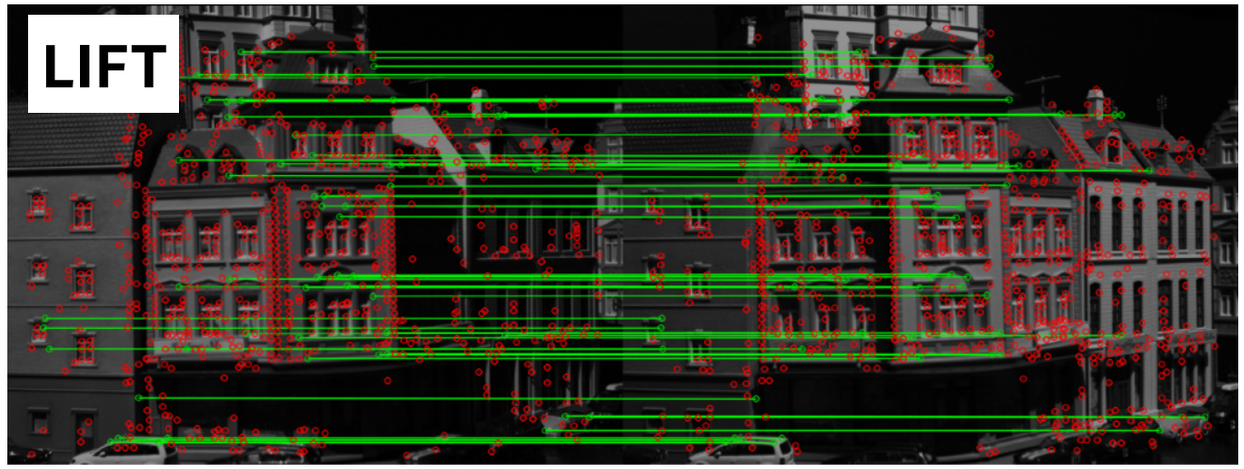
Detected Point <u>Superset</u>

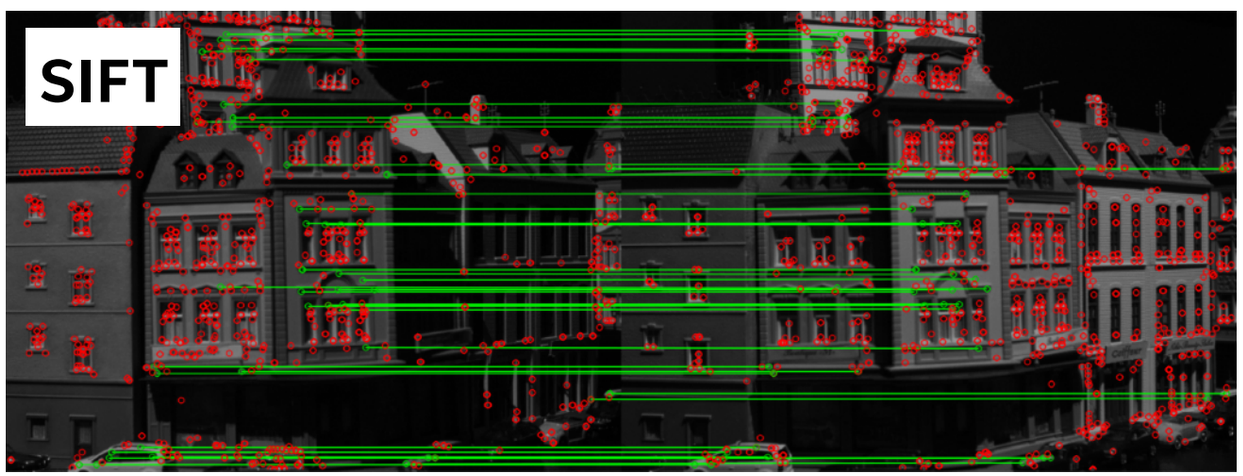


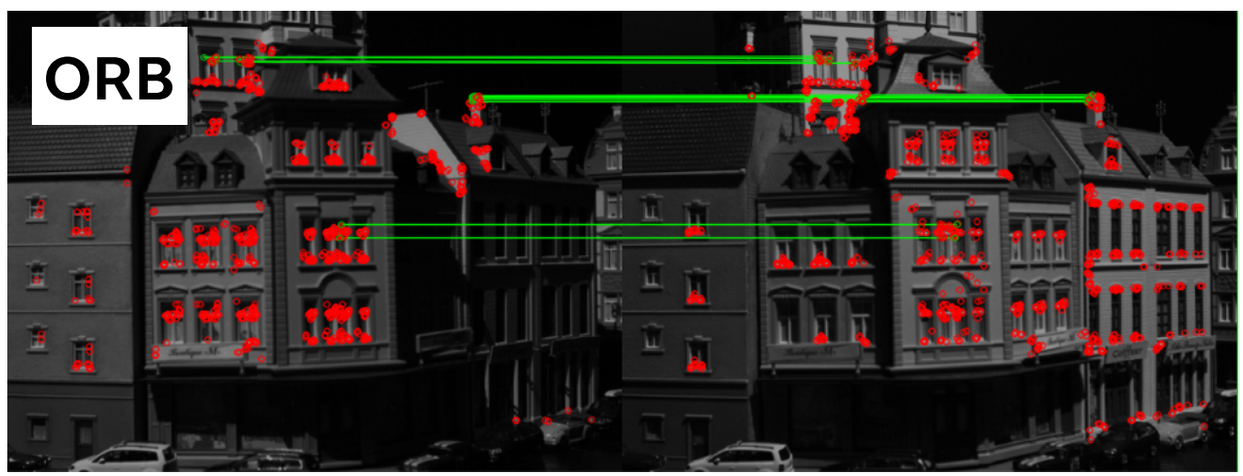
Qualitative Illumination Example

- SuperPoint -> denser set of correct matches
- ORB -> highly clustered matches



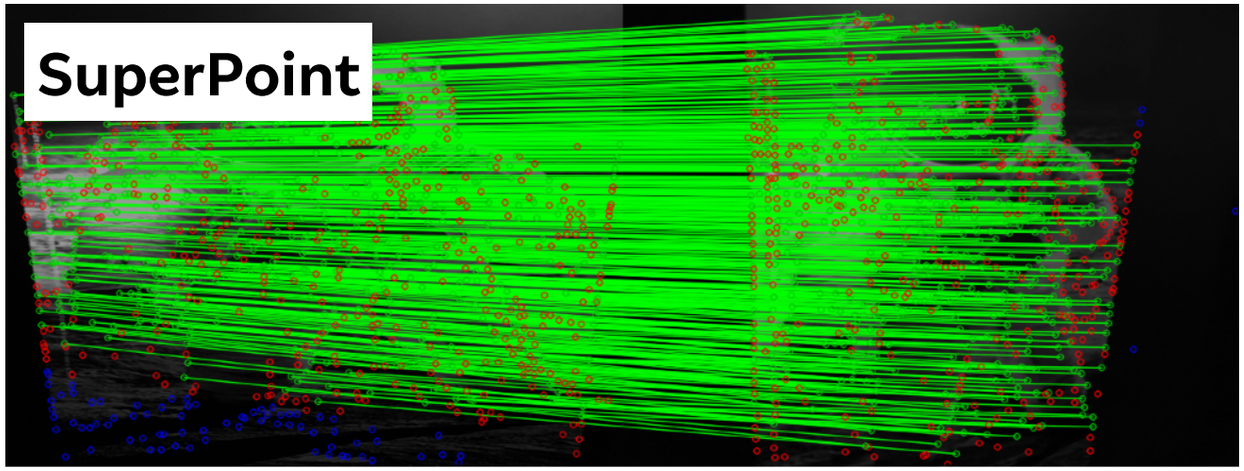


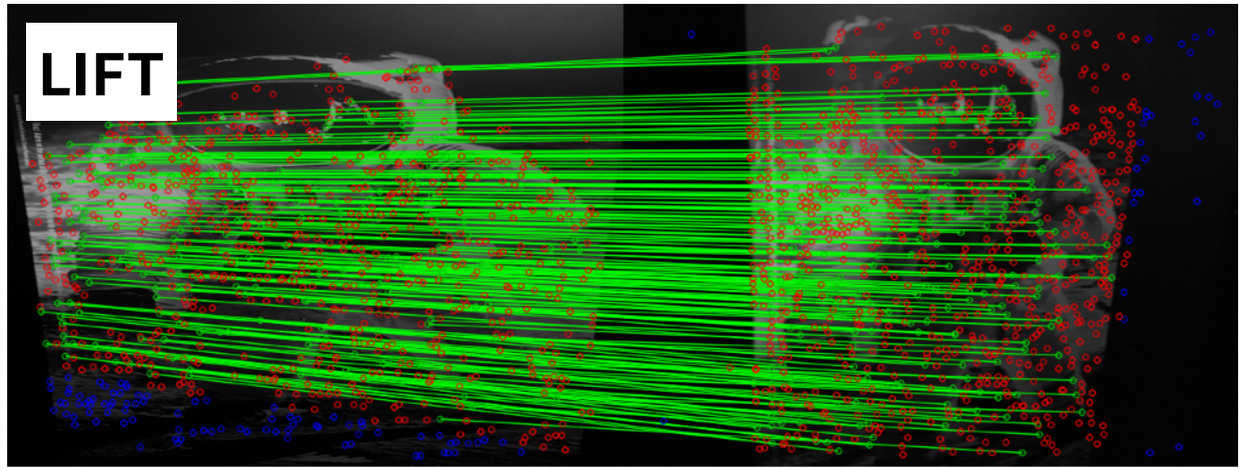


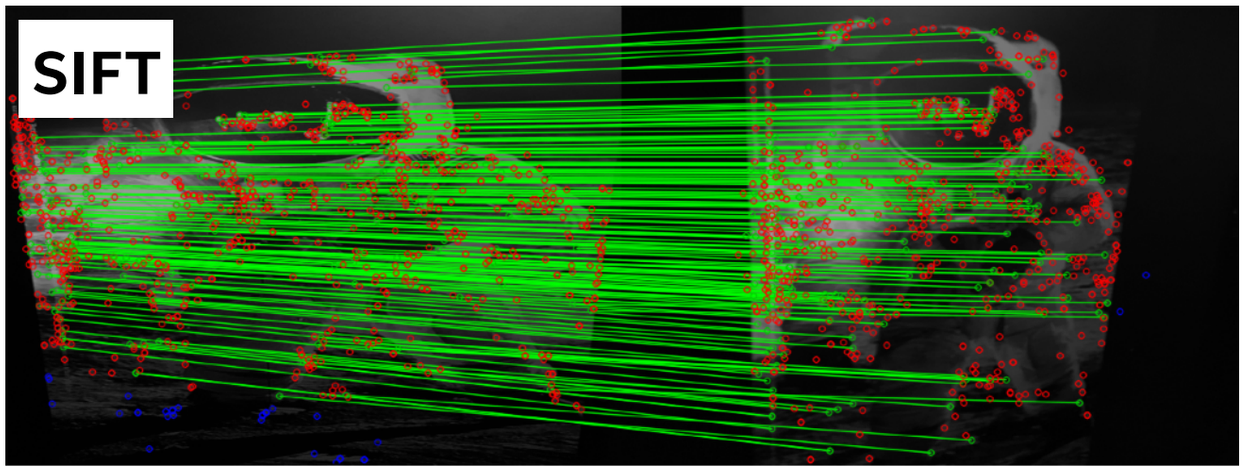


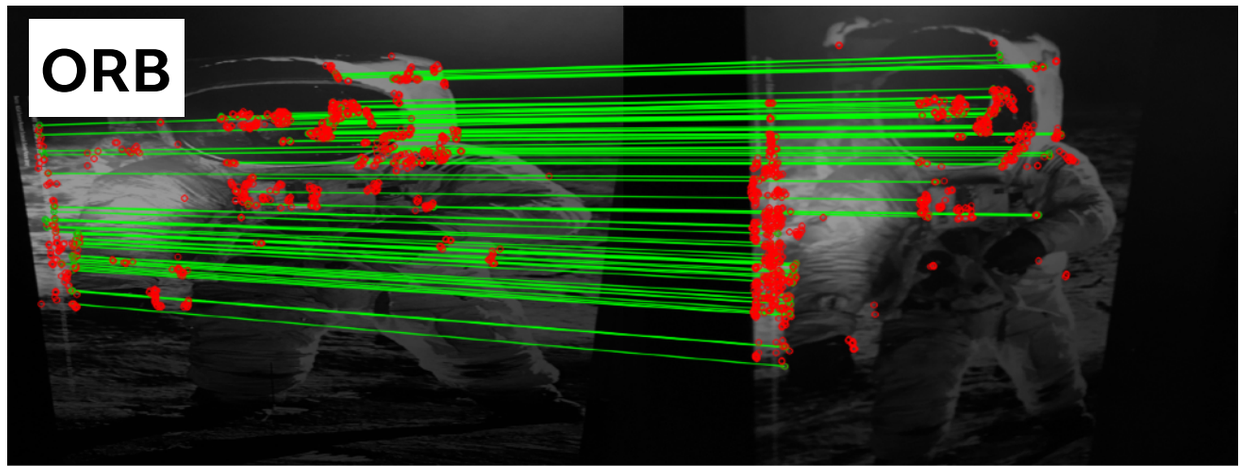
Qualitative Viewpoint Example #1

- Similar story



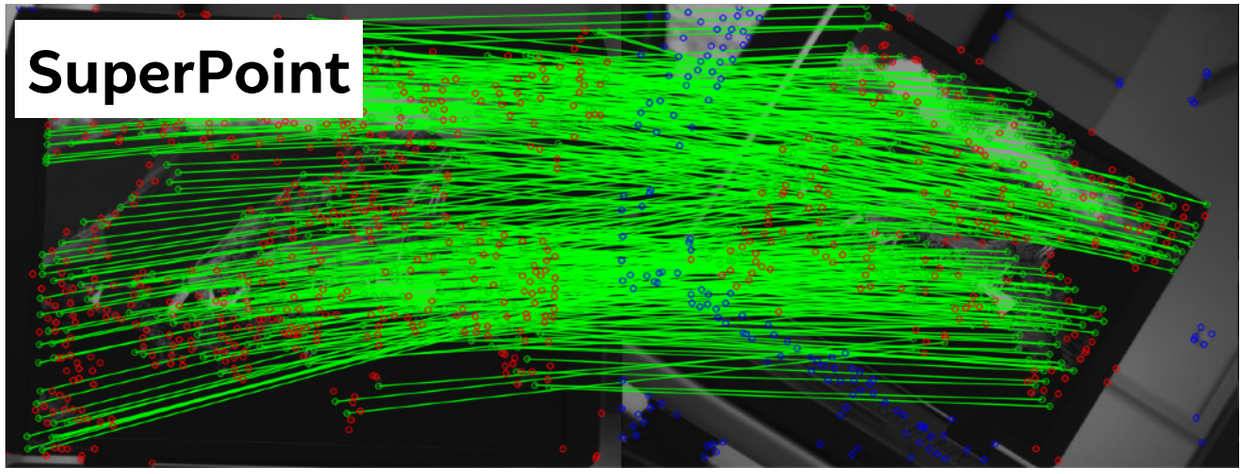


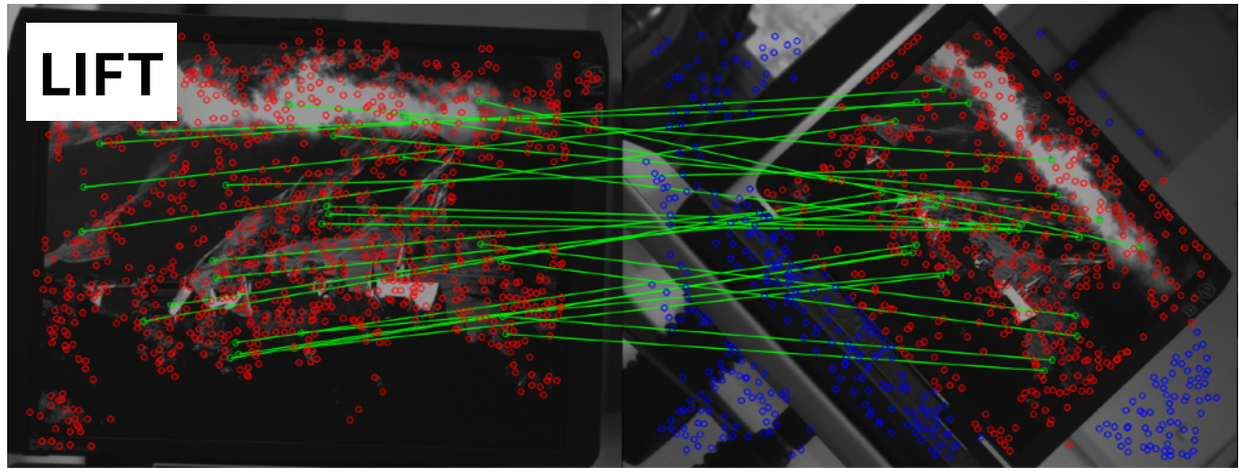


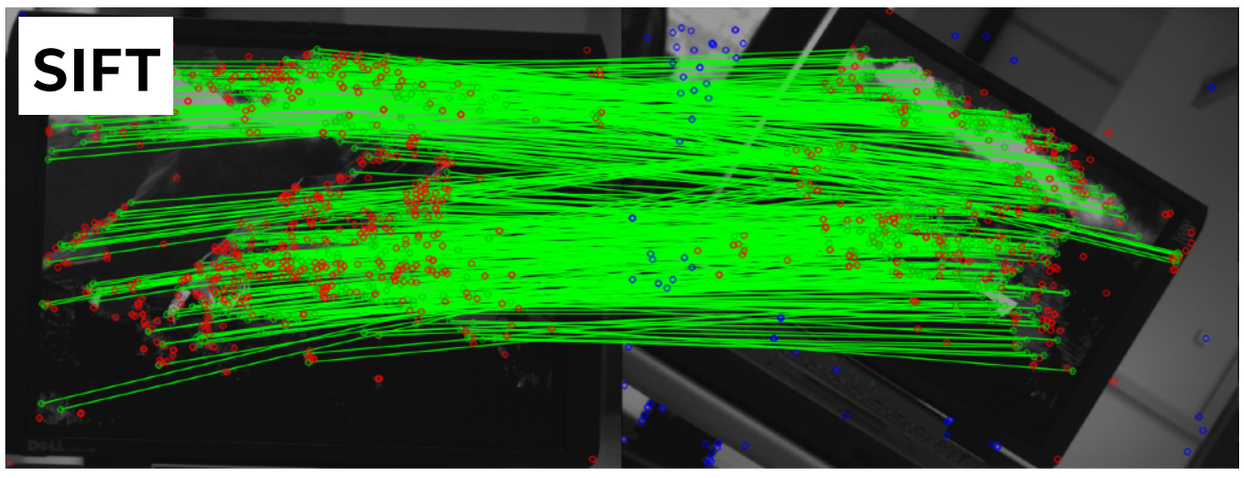


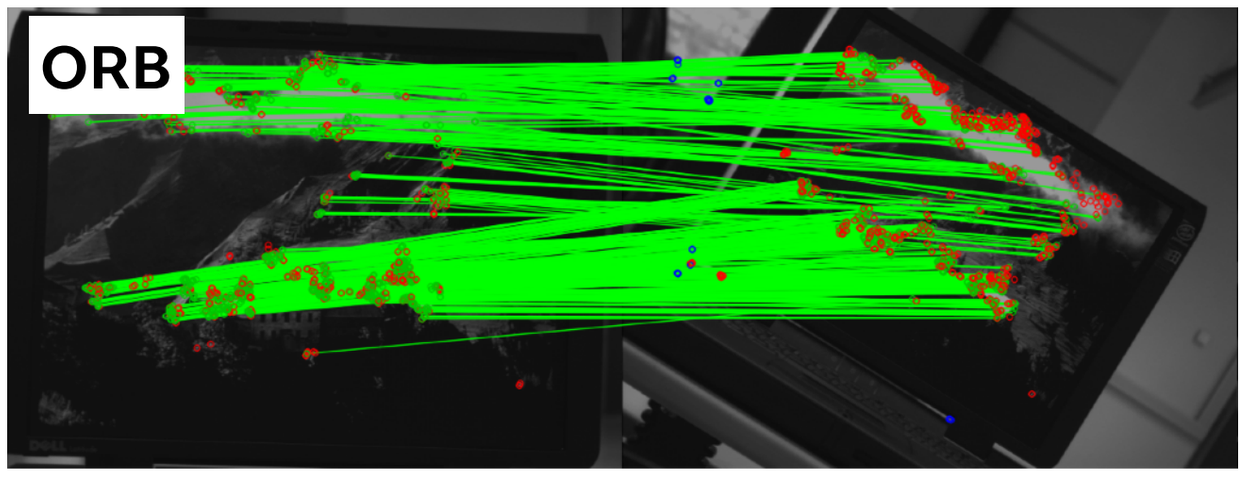
Qualitative Viewpoint Example #2

- In-plane rotation of ~35 degrees









HPatches Evaluation

	Core Task	Sub-metrics			
		Descriptor Metrics		Detector Metrics	
	Homography Estimation ↑	NN mAP ↑	M. Score ↑	Rep. ↑	MLE ↓
SuperPoint	**0.684**	**0.821**	**0.470**	0.581	1.158
LIFT	0.598	0.664	0.315	0.449	1.102
SIFT	0.676	0.694	0.313	0.495	**0.833**
ORB	0.395	0.735	0.266	**0.641**	1.157

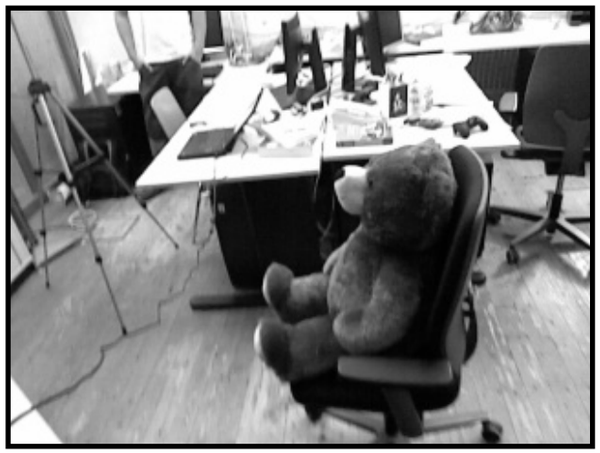
Timing SuperPoint vs LIFT

- Speed important for low-compute Visual SLAM
- SuperPoint total 640x480 time: ~ 33 ms
- LIFT total 640x480 time: ~2 minutes

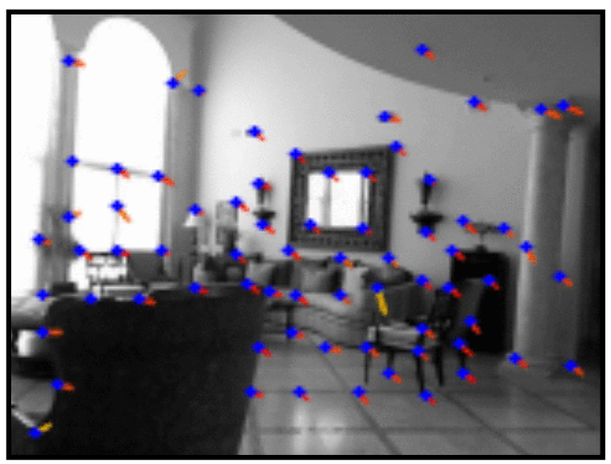
3D Generalizability of SuperPoint

- Trained+evaluated on planar, does it generalize to 3D?
- “Connect-the-dots” using nearest neighbor matches
- Works across many datasets / input modalities / resolutions!

Freiburg (Kinect)

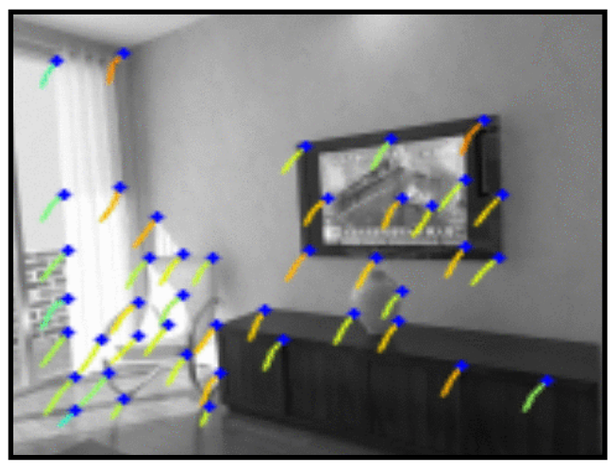


NYU (Kinect)



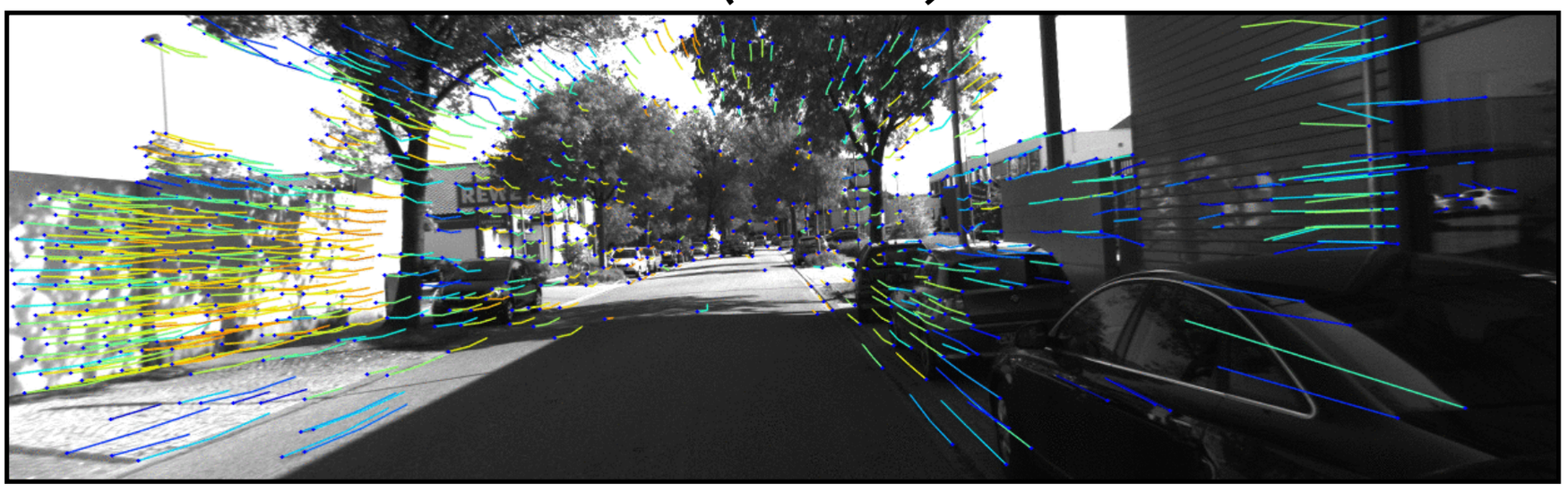
MonoVO (fisheye)

ICL-NUIM (synth)



MS7 (Kinect)

KITTI (stereo)



Download SuperPoint from Research @ MagicLeap

Public Release of Pre-trained Net:

github.com/MagicLeapResearch/SuperPointPretrainedNetwork

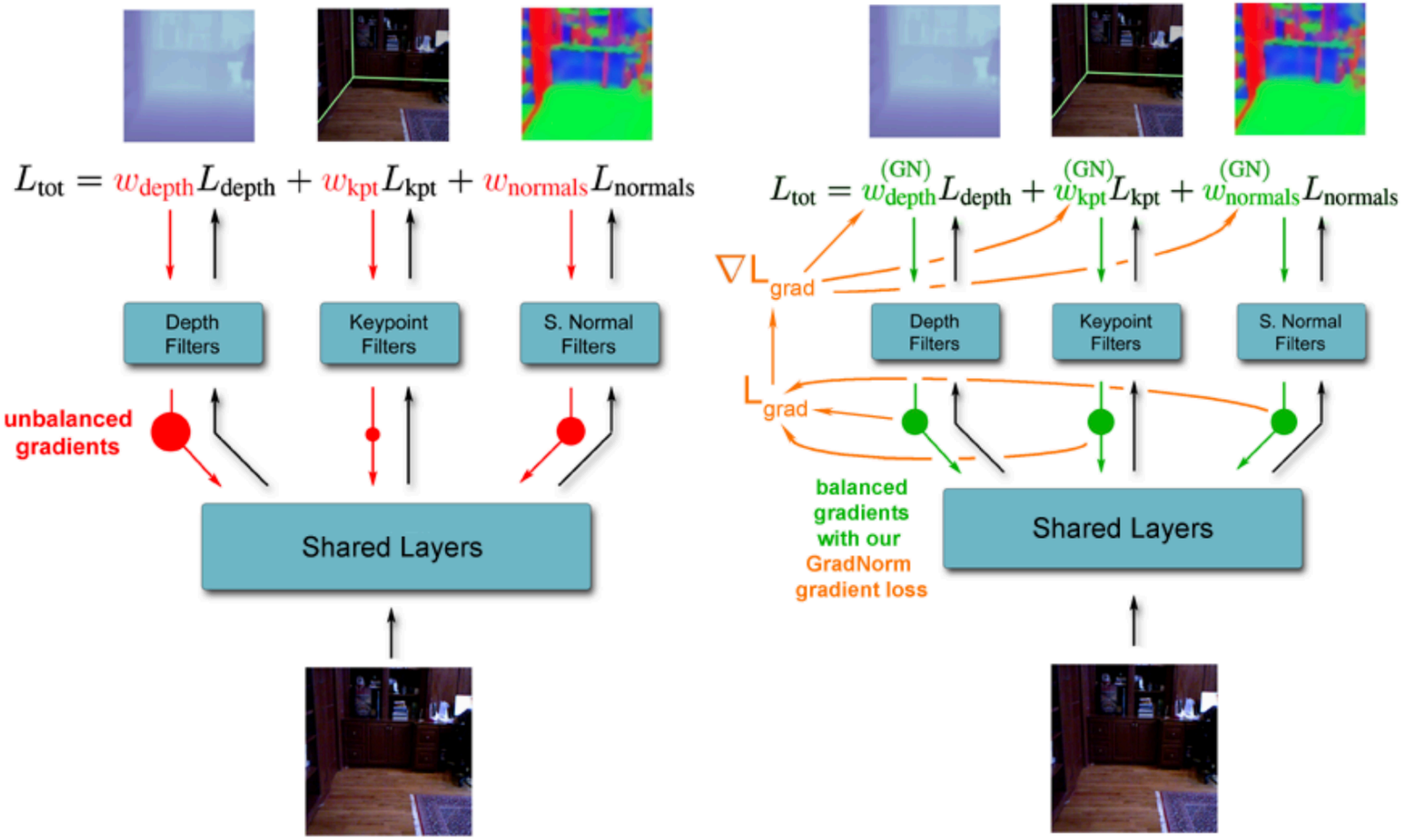
- Sparse Optical Flow Tracker Demo
- Implemented in Python + PyTorch
- Two files, minimal dependencies
- Easy to get up and running

SuperPoint Take-Aways

- “SuperPoint”: A Modern Deep SLAM Frontend
  - Fully convolutional network for real-time deployability
  - Self-supervised recipe to train keypoints
    - Synthetic pre-training + Homography-based adaptation
  - Public code available to run SuperPoint
    - Get up and running in 5 minutes, or your money back

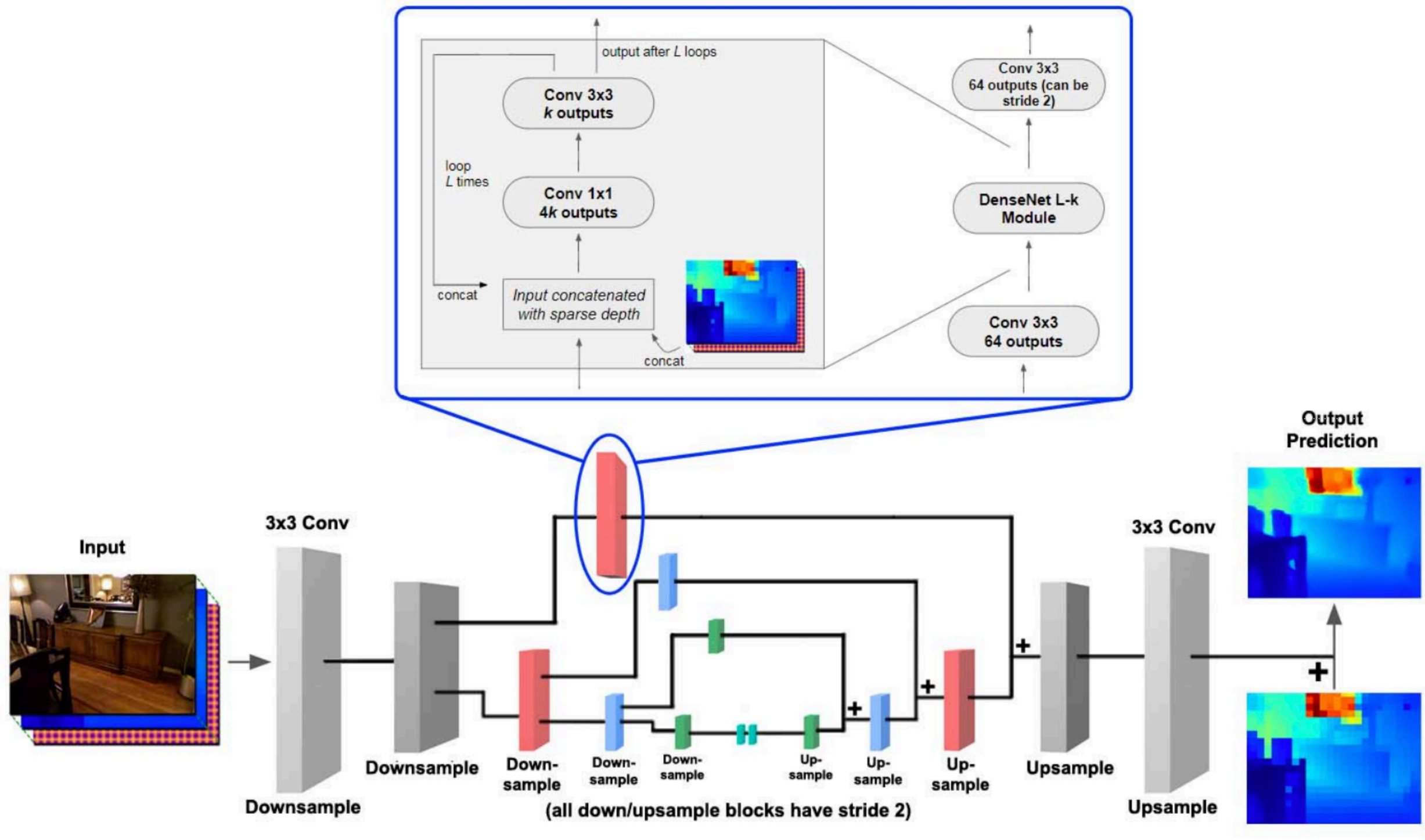
DeTone et al. **SuperPoint: Self-supervised interest point detection and description.** In Workshop on Deep Learning for Visual SLAM at Computer Vision and Pattern Recognition (CVPR), 2018.

GradNorm for Multi-task Learning



Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In ICML, 2018.

Deep Depth Densification



Chen, Z., Badrinarayanan, V., Drozdov, G., Rabinovich, A. Estimating Depth from RGB and Sparse Sensing. In ECCV, 2018.

Z. Chen et al. (ECCV 2018) Estimating Depth from RGB and Sparse Sensing

RGB

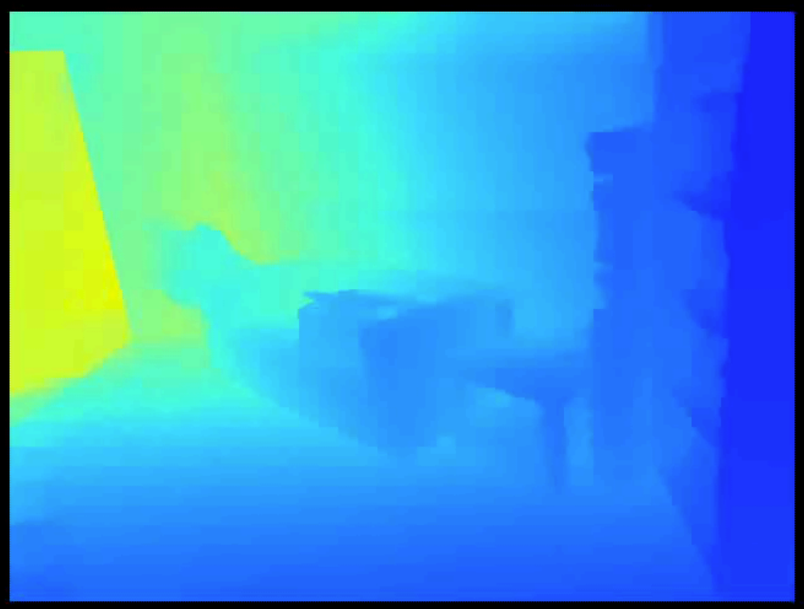
Sparse Depth (Overlaid)

0.00%/100.00% Points Sampled

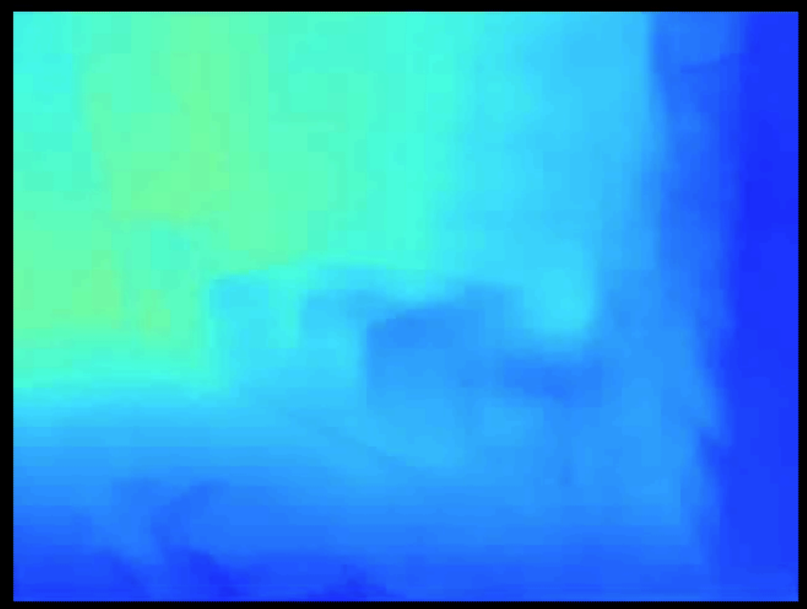
RMS Error, Current Image: 0.542m

Increasing Sparse Samples in 8s

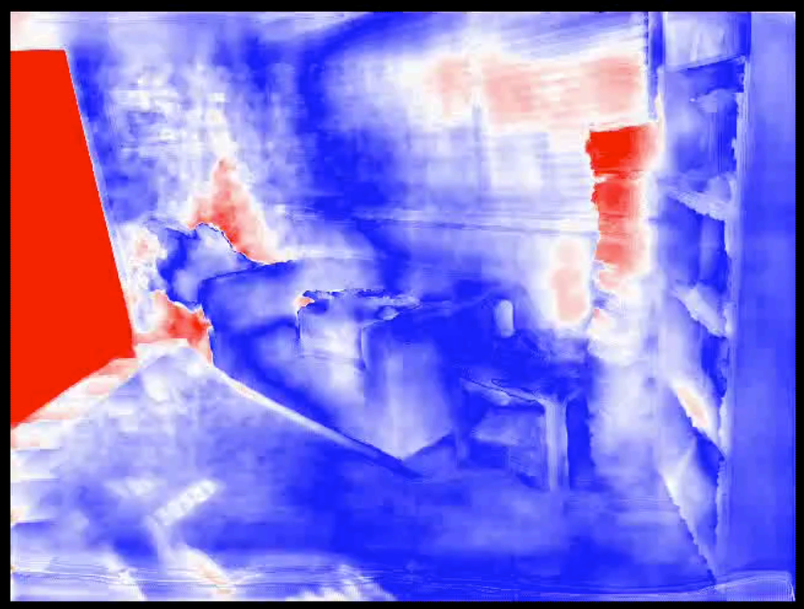
Ground Truth



Our D3 Output



Error Map



Looking for research interns (PhD students)

Looking for hybrid researchers/engineers for full-time roles

eccv@magicleap.com

Location #1: San Francisco Bay Area, California

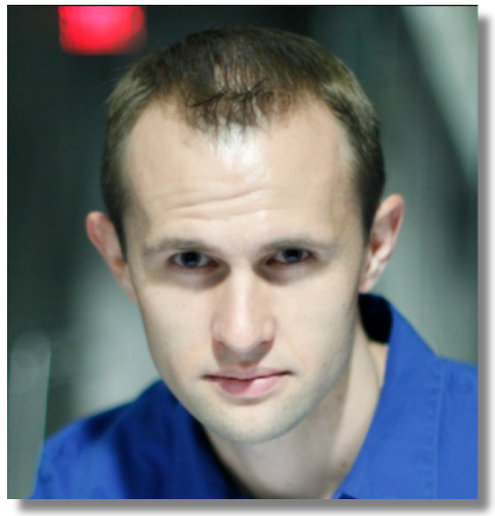
Location #2: Zurich, Switzerland

References

Daniel DeTone

Zhao (Joe) Chen

Vijay Bardinarayan



Tomasz Malisiewicz

Andrew Rabinovich

DeTone, D., Malisiewicz, T., Rabinovich, A. Toward Geometric DeepSLAM. In arXiv: 1707.07410.

DeTone, D., Malisiewicz, T., Rabinovich, A. SuperPoint: Self-supervised interest point detection and description. In Workshop on Deep Learning for Visual SLAM at Computer Vision and Pattern Recognition (CVPR), 2018.

Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In ICML, 2018.

Chen, Z., Badrinarayanan, V., Drozdov, G., Rabinovich, A. Estimating Depth from RGB and Sparse Sensing. In ECCV, 2018.

Thank You

magic leap

SuperPoint: A Modern Deep SLAM Front-end