SuperPoint: Self-Supervised Interest Point Detection and Description

Daniel DeTone, Tomasz Malisiewicz, Andrew Rabinovich

(CVPR 2018)

10.18.2022

전북대학교 학부생 김세희

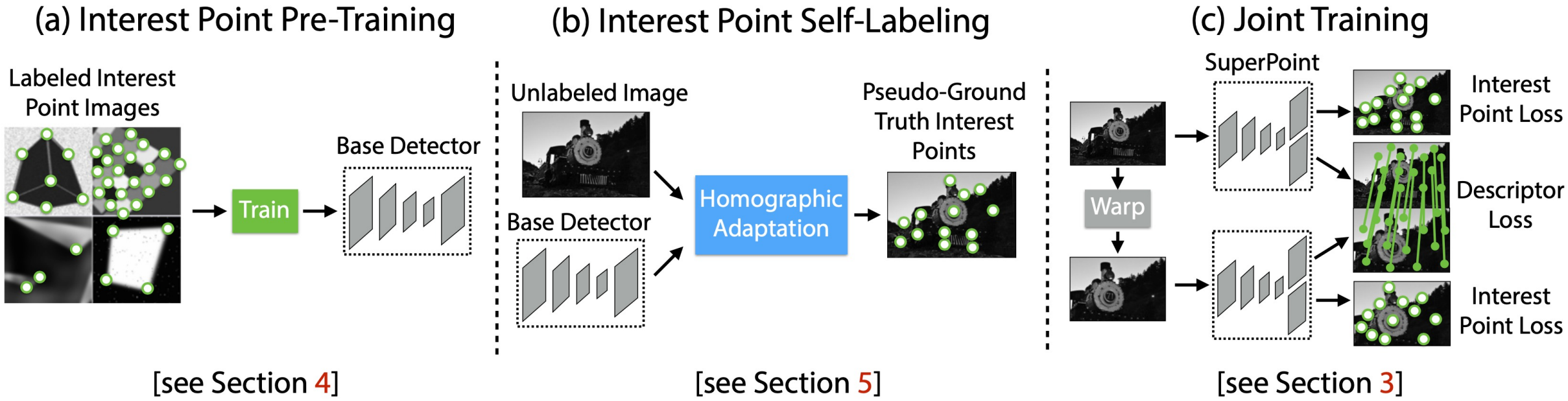
Introduction

- Convolutional neural networks have been shown to be superior to hand-engineered representations on almost all tasks requiring images as input.
- At the heart of these techniques is a large dataset of 2D ground truth locations labeled by human annotators.
- Instead of using human supervision to define interest points in real images, we present a self-supervised solution using self-training
- In our approach, we create a large dataset of pseudo-ground truth interest point locations in real images, supervised by the interest point detector itself, rather than a large-scale human annotation effort.

Method

Self-Supervised Training Overview

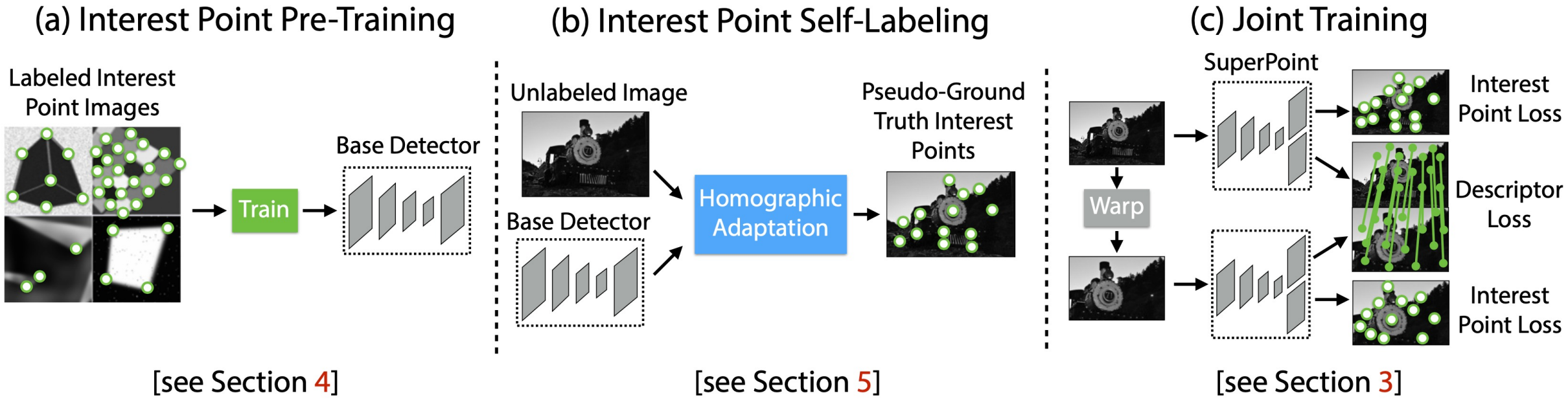
- To generate the pseudo-ground truth interest points, we first train a fully-convolutional neural network on millions of examples from a synthetic dataset we created called Synthetic Shapes (see Figure 2a).



Method

Self-Supervised Training Overview

- We call the resulting trained detector *MagicPoint* - it significantly outperforms traditional interest point detectors on the synthetic dataset (see Section 4).



Method

Synthetic Pre-Training

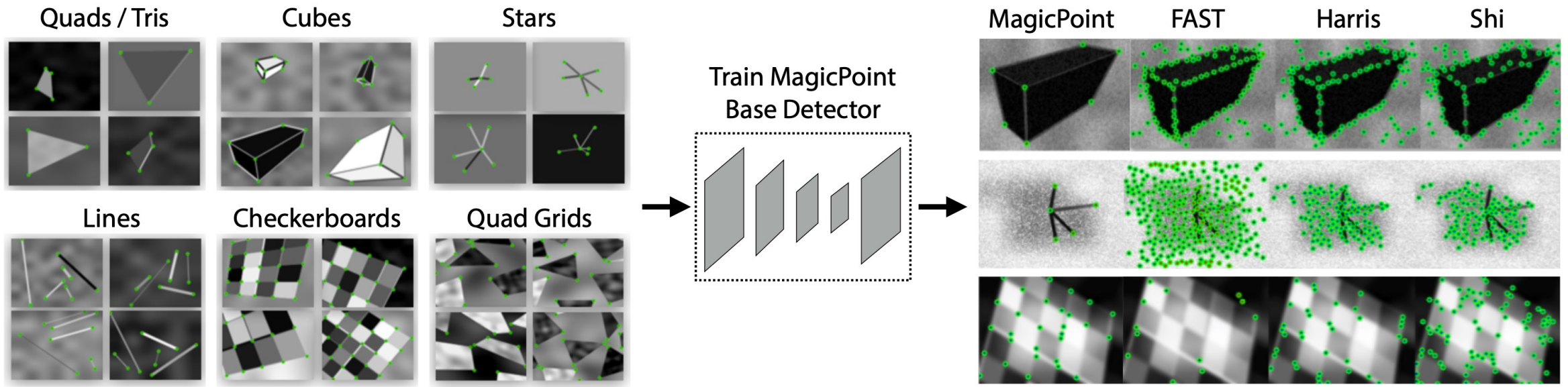
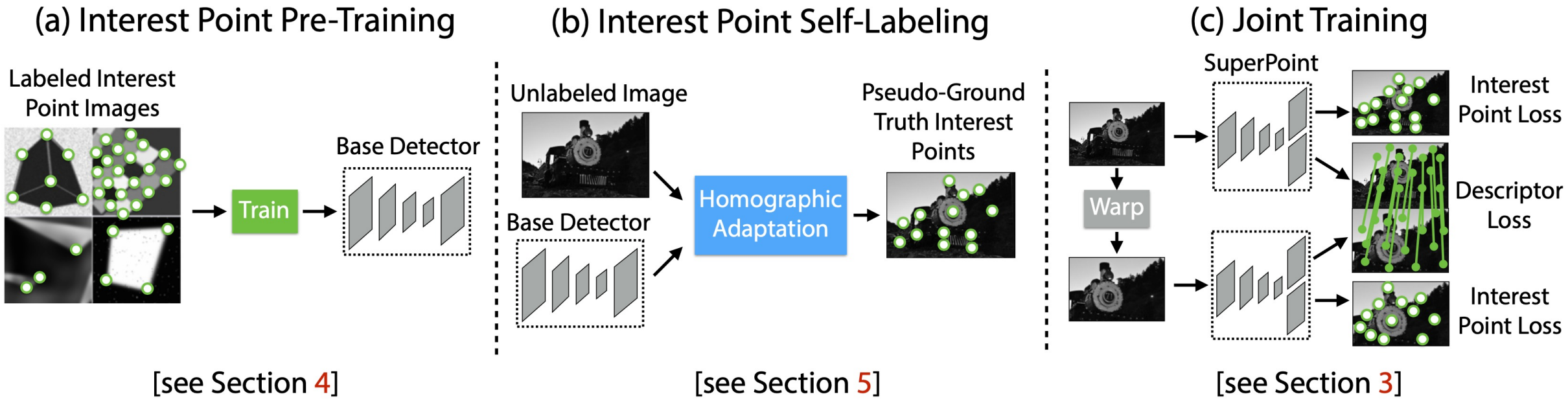


Figure 4. **Synthetic Pre-Training.** We use our Synthetic Shapes dataset consisting of rendered triangles, quadrilaterals, lines, cubes, checkerboards, and stars each with ground truth corner locations. The dataset is used to train the MagicPoint convolutional neural network, which is more robust to noise when compared to classical detectors.

Method

Self-Supervised Training Overview

- To bridge this gap in performance on real images, we developed a multi-scale, multi-transform technique – *Homographic Adaptation*.



Method

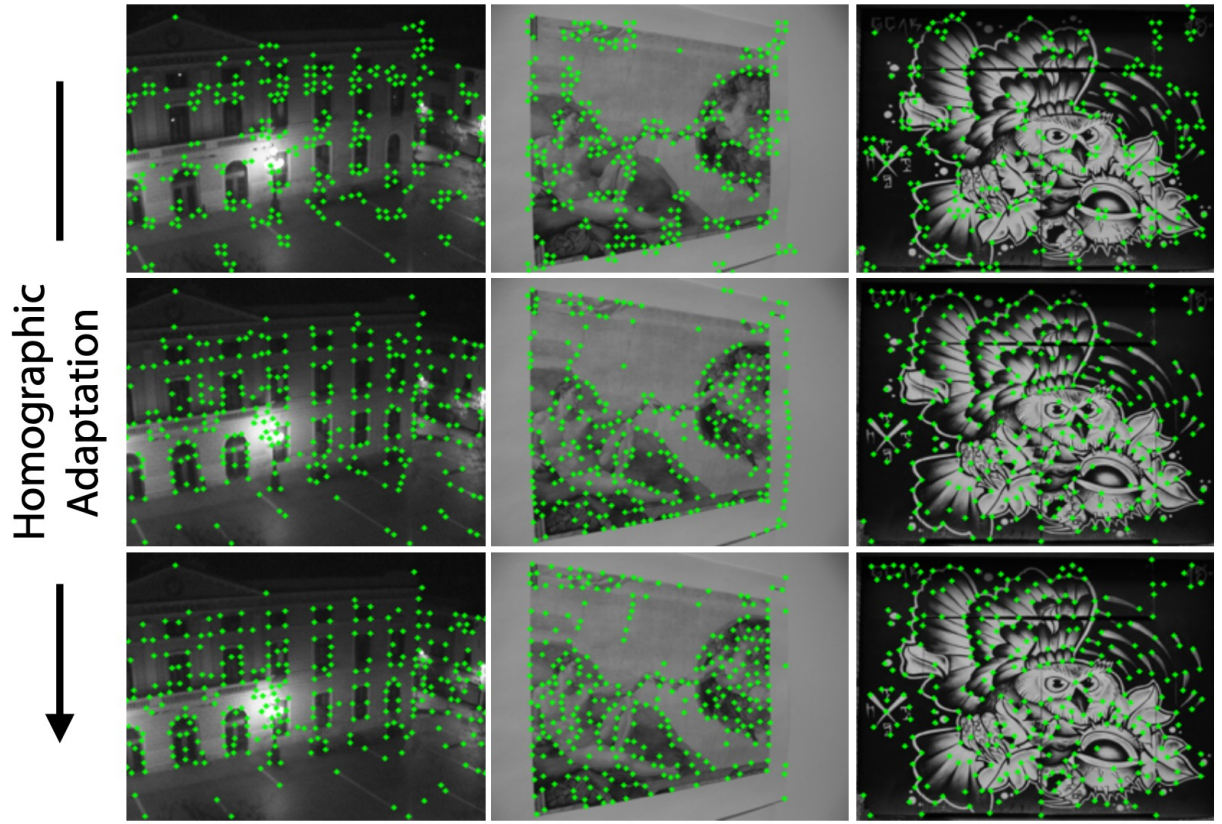
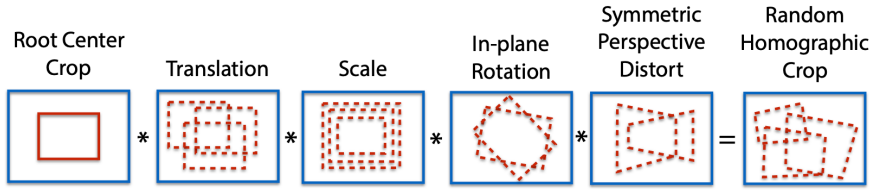


Figure 7. **Iterative Homographic Adaptation**. Top row: initial base detector (MagicPoint) struggles to find repeatable detections. Middle and bottom rows: further training with Homographic Adaption improves detector performance.

Method

Homographic Adaptation



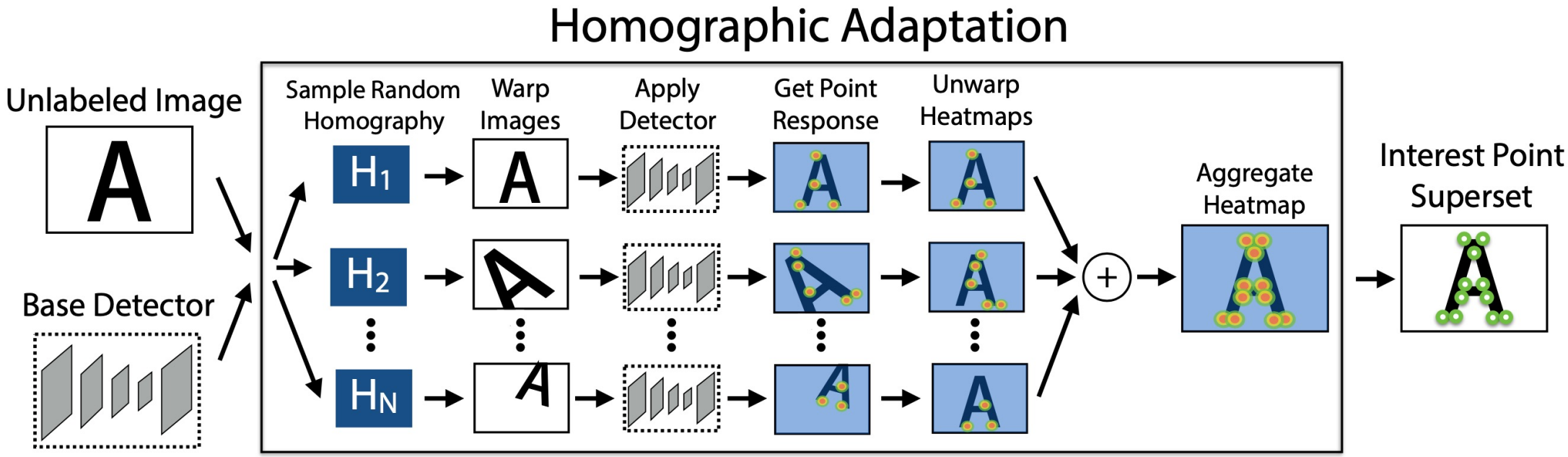
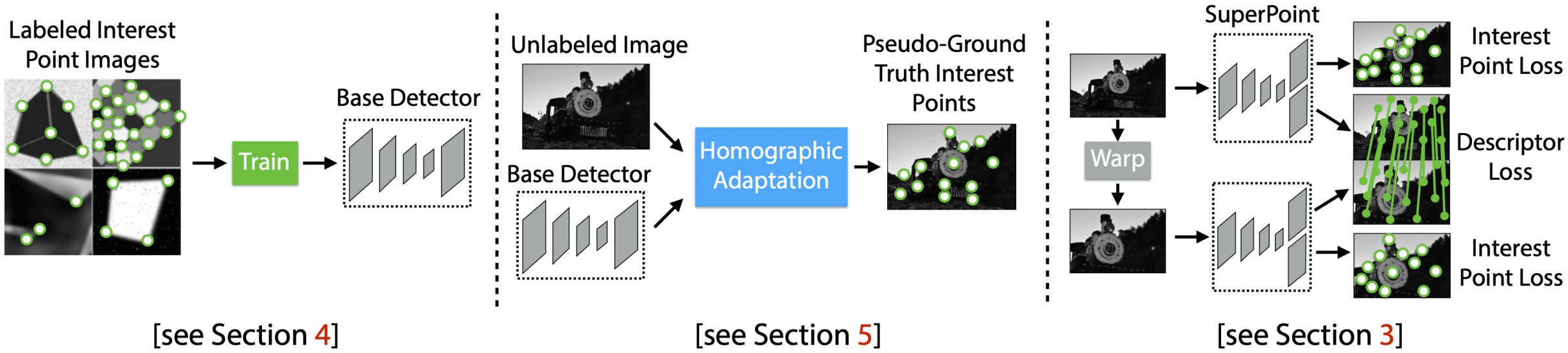


Figure 5. **Homographic Adaptation.** Homographic Adaptation is a form of self-supervision for boosting the geometric consistency of an interest point detector trained with convolutional neural networks. The entire procedure is mathematically defined in Equation 10.

Method

Self-Supervised Training Overview

- We lastly combine SuperPoint with a descriptor subnetwork (see Figure 2c).



Method

The Resulting System

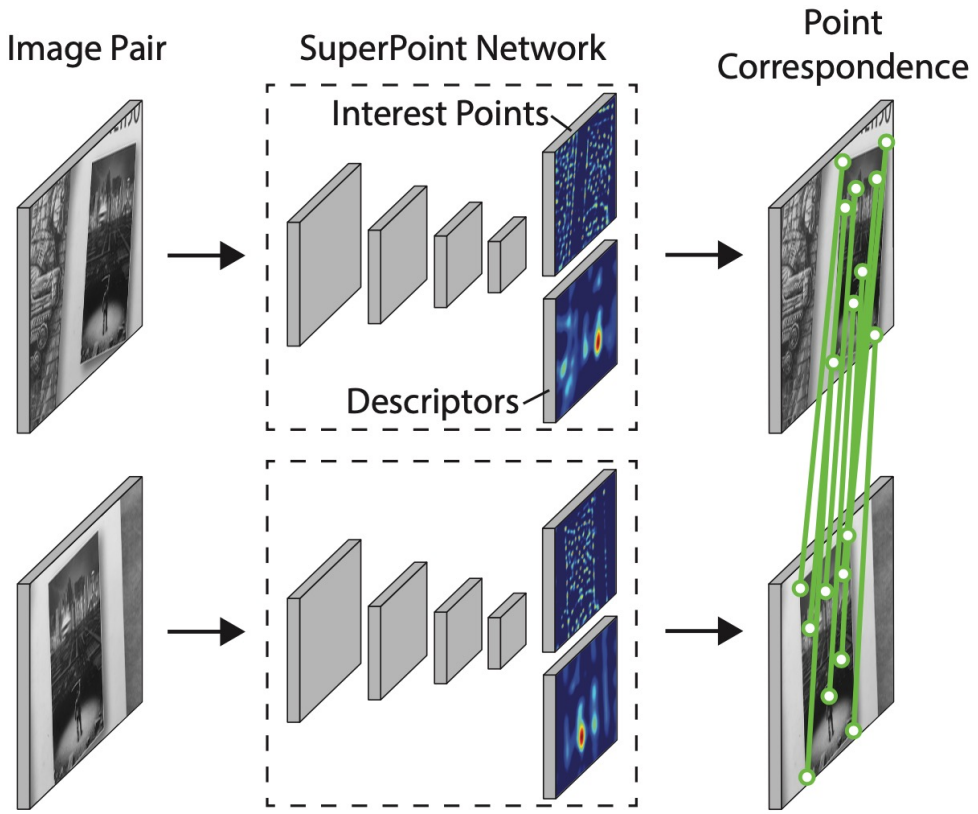


Figure 1. **SuperPoint for Geometric Correspondences.** We present a fully-convolutional neural network that computes SIFT-like 2D interest point locations and descriptors in a single forward pass and runs at 70 FPS on 480×640 images with a Titan X GPU.

Method

Shared Encoder

- Our SuperPoint architecture uses a VGG-style [27] encoder to reduce the dimensionality of the image. The encoder consists of convolutional layers.

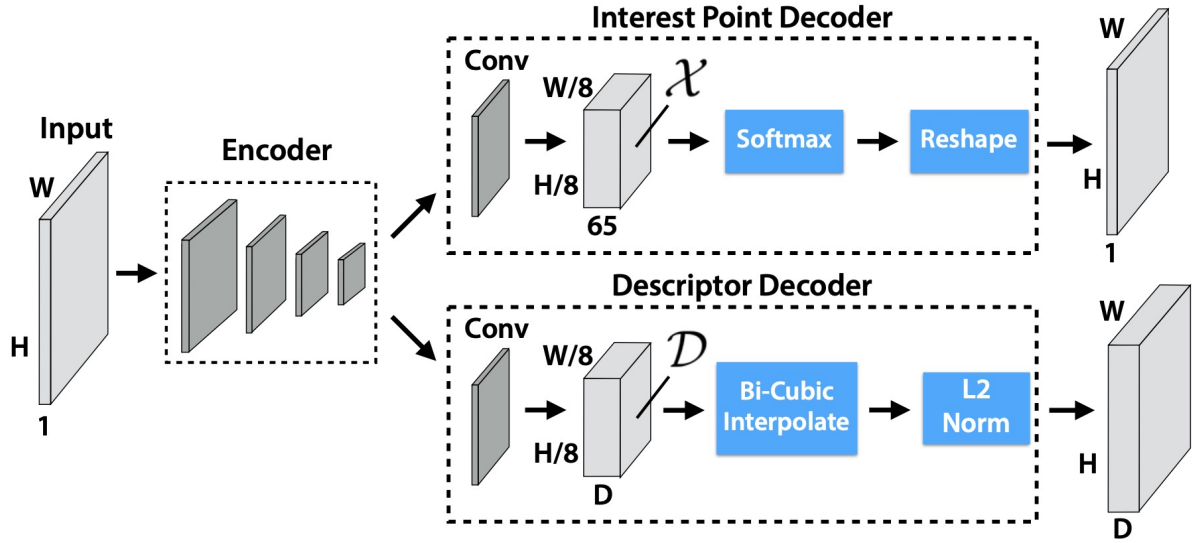


Figure 3. **SuperPoint Decoders**. Both decoders operate on a shared and spatially reduced representation of the input. To keep the model fast and easy to train, both decoders use non-learned upsampling to bring the representation back to $\mathbb{R}^{H \times W}$.

Experiments

	57 Illumination Scenes		59 Viewpoint Scenes	
	NMS=4	NMS=8	NMS=4	NMS=8
SuperPoint	**.652**	**.631**	.503	**.484**
MagicPoint	.575	.507	.322	.260
FAST	.575	.472	.503	.404
Harris	.620	.533	**.556**	.461
Shi	.606	.511	.552	.453
Random	.101	.103	.100	.104

Table 3. **HPatches Detector Repeatability**. SuperPoint is the most repeatable under illumination changes, competitive on viewpoint changes, and outperforms MagicPoint in all scenarios.

Experiments

	Homography Estimation			Detector Metrics		Descriptor Metrics	
	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 5$	Rep.	MLE	NN mAP	M. Score
SuperPoint	.310	**.684**	**.829**	.581	1.158	**.821**	**.470**
LIFT	.284	.598	.717	.449	1.102	.664	.315
SIFT	**.424**	.676	.759	.495	**0.833**	.694	.313
ORB	.150	.395	.538	**.641**	1.157	.735	.266

Table 4. **HPatches Homography Estimation.** SuperPoint outperforms LIFT and ORB and performs comparably to SIFT using various ϵ thresholds of correctness. We also report related metrics which measure detector and descriptor performance individually.

Experiments

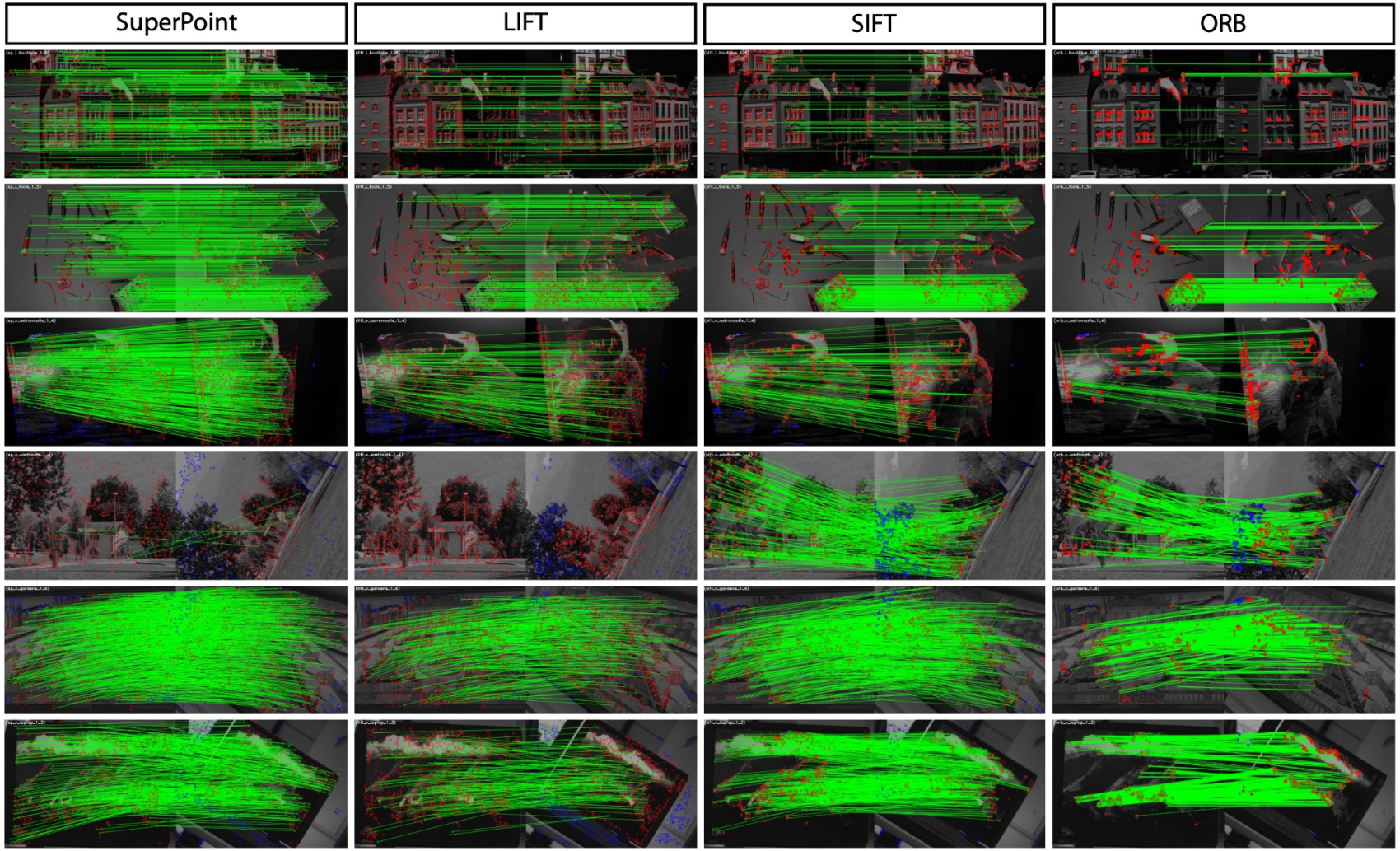


Figure 8. **Qualitative Results on HPatches**. The green lines show correct correspondences. SuperPoint tends to produce more dense and correct matches compared to LIFT, SIFT and ORB. While ORB has the highest average repeatability, the detections cluster together and generally do not result in more matches or more accurate homography estimates (see 4). Row 4: Failure case of SuperPoint and LIFT due to extreme in-plane rotation not seen in the training examples. See Appendix D for additional homography estimation example pairs.

Thank you