SuperPoint

Friday, December 27, 2019 9:28 PM

→ Fully convolutional n/w that computes pixel-level interest point locations and descriptors.

→ State of the art homography estimation results on HPatches compared to LIFT, SIFT and ORB.

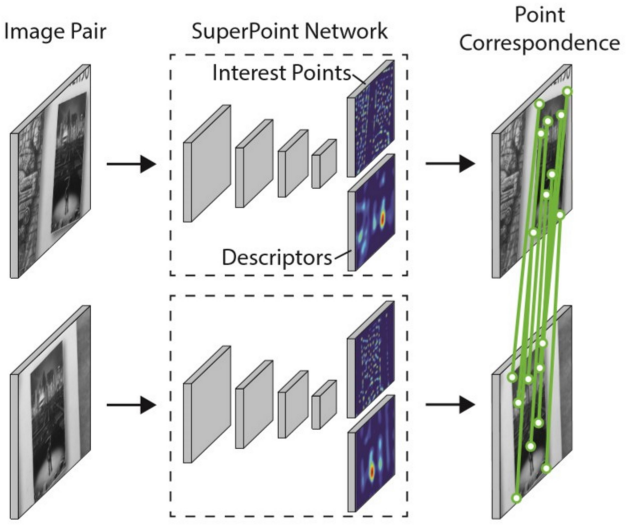

Figure 1. **SuperPoint for Geometric Correspondences.** We present a fully-convolutional neural network that computes SIFT-like 2D interest point locations and descriptors in a single forward pass and runs at 70 FPS on 480×640 images with a Titan X GPU.

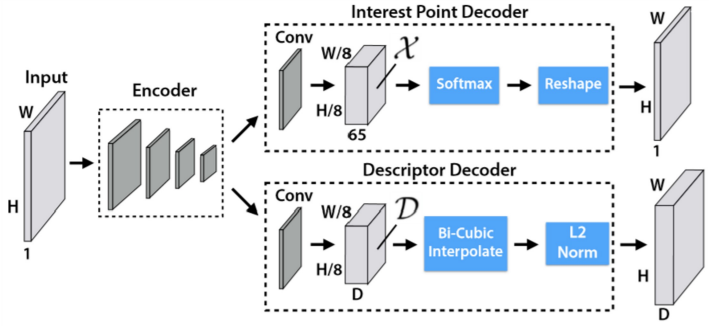

Figure 3. **SuperPoint Decoders**. Both decoders operate on a shared and spatially reduced representation of the input. To keep the model fast and easy to train, both decoders use non-learned upsampling to bring the representation back to $\mathbb{R}^{H \times W}$.

→ One encoder and 2 decoders with shared weights.

→ Encoder → VGG-style with Max-pooling layers.

→ Interest point decoder → uses sub-pixel convolution or depth to space (TF) or Pixel-shuffle (PyTorch).
→ this has much lower computation as compared to traditional deconv. layers.

→ Descriptor → semi-dense, normalized to unit length, Bicubic interpolated to get a dense map.

21

$$\mathcal{L}(\mathcal{X}, \mathcal{X}', \mathcal{D}, \mathcal{D}'; Y, Y', S) = \mathcal{L}_p(\mathcal{X}, Y) + \mathcal{L}_p(\mathcal{X}', Y') + \lambda \mathcal{L}_d(\mathcal{D}, \mathcal{D}', S). \quad (1)$$

$\mathcal{L}_p$ → Cross entropy loss on [illegible] interest point

- cross entropy loss
  Pseudo-GT interest point locations

L_D → Descriptor loss b/w all pairs of descriptors b/w 2 images

→ Synthetic pre-training
Synthetic Shapes dataset with quadrilaterals, Δles, lines and ellipses. Use L, T and Y junctions as interest points.

→ Magic point → Only the detector pathway.
This worked extremely well on synthetic images and generalized reasonably to images with strong corners.
However didn't work well

However didn't work well on natural images.

Hence a self-supervised approach was used called Homography Adaptation.

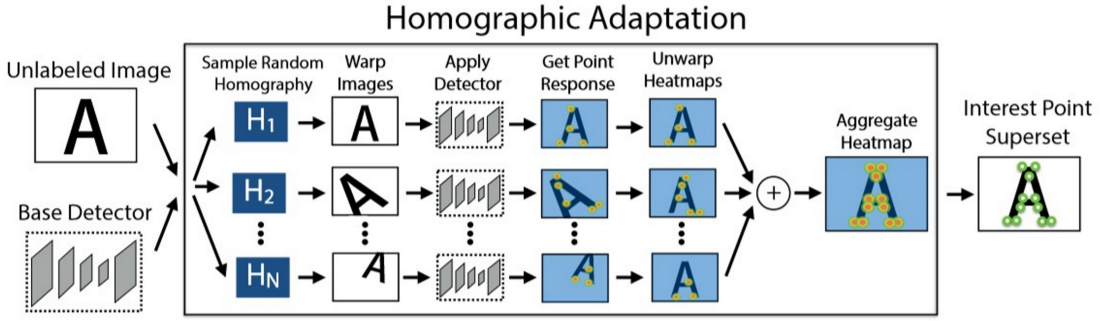

Figure 5. **Homographic Adaptation.** Homographic Adaptation is a form of self-supervision for boosting the geometric consistency of an interest point detector trained with convolutional neural networks. The entire procedure is mathematically defined in Equation 10.

$$\mathbf{x} = \mathcal{H}^{-1} f_\theta(\mathcal{H}(I)).$$

- The interest point should be covariant wrt transform
- An aggregation over a large number of H's gives rise to SuperPoint

$$\hat{F}(I; f_\theta) = \frac{1}{N_h} \sum_{i=1}^{N_h} \mathcal{H}_i^{-1} f_\theta(\mathcal{H}_i(I)). \qquad (10)$$

→ To have plausible camera transformation H is decomposed into translation, scale, rotation (Yaw) and symmetric distortion.

→ 100 homographies per image are used.

→ Architecture →

VGG like 8 Conv layers.

Encoder →

(3×3) all $C_{64} \rightarrow C_{64} \rightarrow M_{\downarrow 2} \rightarrow C_{64} \rightarrow C_{64} \rightarrow M_{\downarrow 2}$

$\downarrow$

$C_{128} \leftarrow M_{\downarrow 2} \leftarrow C_{128} \leftarrow C_{128}$

$\downarrow$

$C_{128} \rightarrow M_{\downarrow 2}$

Decoder →

C_{256} (3×3) $\rightarrow$ $C_{65 \text{ or } 256}$ (1×1)

C → Conv → ReLU → BN

Trained on synthetic shapes.

→ Magic point (Base detector) is

Trained on synthetic shapes.
Then Magicpoint (prev. o/p) is then trained with descriptor head on MS-COCO with Homography adaptation (100 H per image).
Homography adaptation is done again.

Data is augmented using gaussian noise, motion blur and brightness level changes.

→ For a 640 x 480 image to sample 1000 feature points, SuperPoint takes ~13 ms on a TitanXp. (70 FPS).