Dense Segmentation-aware Descriptors

Eduard Trulls[1], Iasonas Kokkinos[2], Alberto Sanfeliu[1], Francesc Moreno-Noguer[1]
[1] Institut de Robòtica i Informàtica Industrial, Barcelona, Spain / [2] Center for Visual Computing, Ecole Centrale de Paris/INRIA-Saclay, France

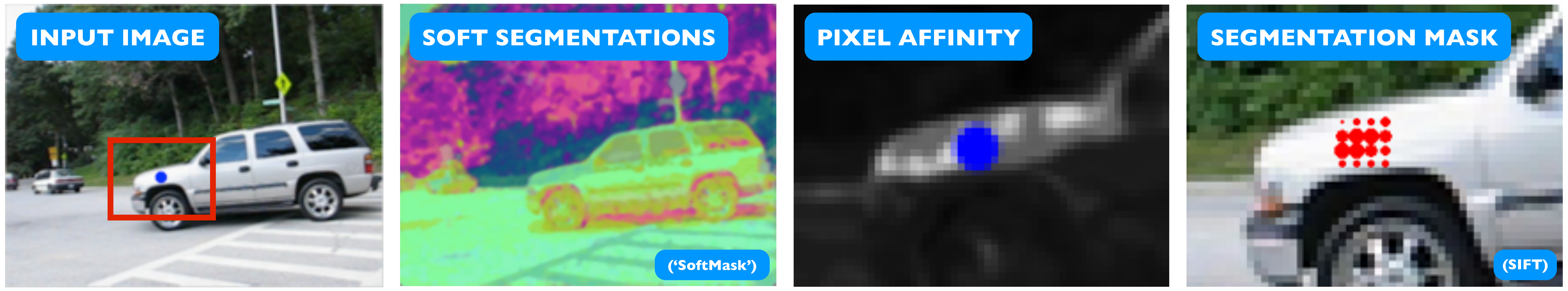
CONTRIBUTIONS

We use **soft segmentation** to **suppress background** structures during **descriptor construction**.
Improvements in **motion** and **stereo**, using both **SID** [1] and **SIFT**.

KEY FEATURES

- **General:** two descriptors, two soft segmentations, two problems.
- **Low-level**: application-independent, no training necessary.
- **Small overhead:** a few seconds.
- **Single parameter:** fixed once, used throughout experiments.

Code: http://www.iri.upc.edu/people/etrulls/#code

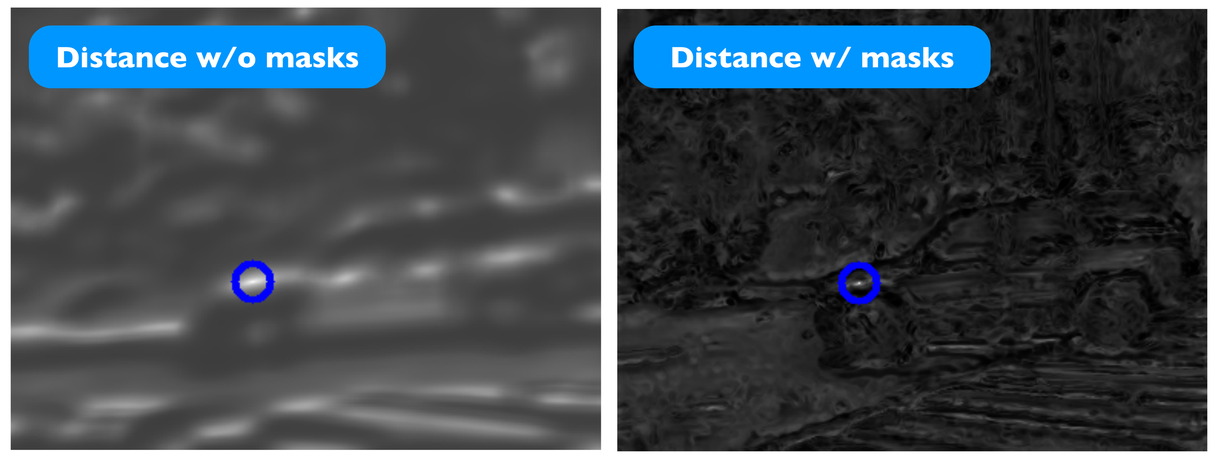


MASK COMPUTATION

Point $\mathbf{x}$, pixel embedding $y(\cdot)$, grid $\mathbf{G}^{[i]}(\mathbf{x}), i = 1 \dots K$ and design parameter λ:

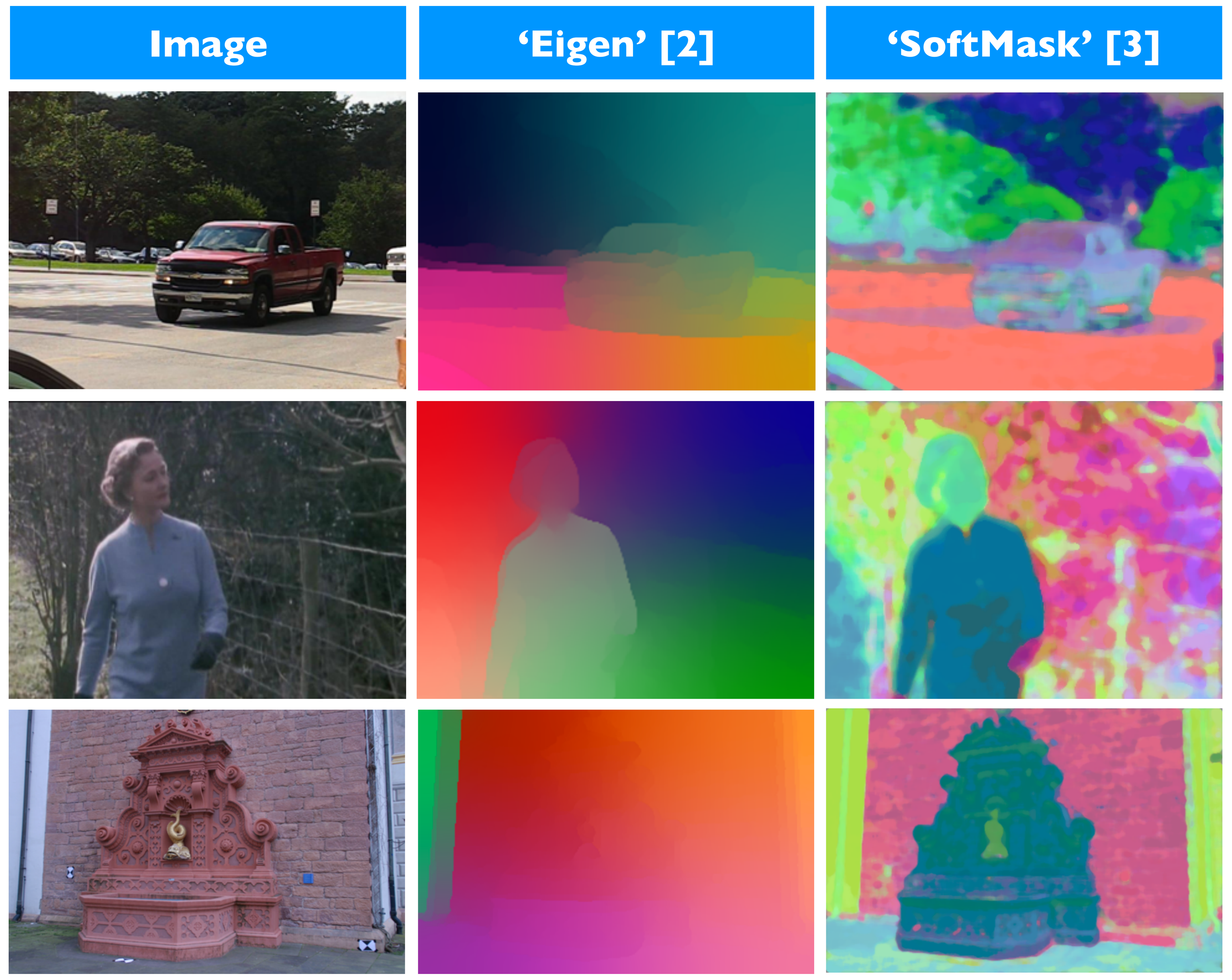
Affinity: $d(\mathbf{x}, \mathbf{G}^{[i]}(\mathbf{x})) = \|y(\mathbf{x}) - y(\mathbf{G}^{[i]}(\mathbf{x})\|_2^2$. **Mask**: $\mathbf{w}^{[i]} = \exp\left(-\lambda \cdot d(\mathbf{x}, \mathbf{G}^{[i]}(\mathbf{x}))\right)$

Segmentation-aware descriptor: $\mathbf{D}'^{[i]} = \mathbf{w}^{[i]}\mathbf{D}^{[i]}$



SOFT SEGMENTATIONS (PIXEL EMBEDDINGS)

Root of all evil: descriptor's support straddles different objects.
Ideal remedy: constrain descriptor to lie on a **single object**.
Practical solution: **use soft segmentations**.

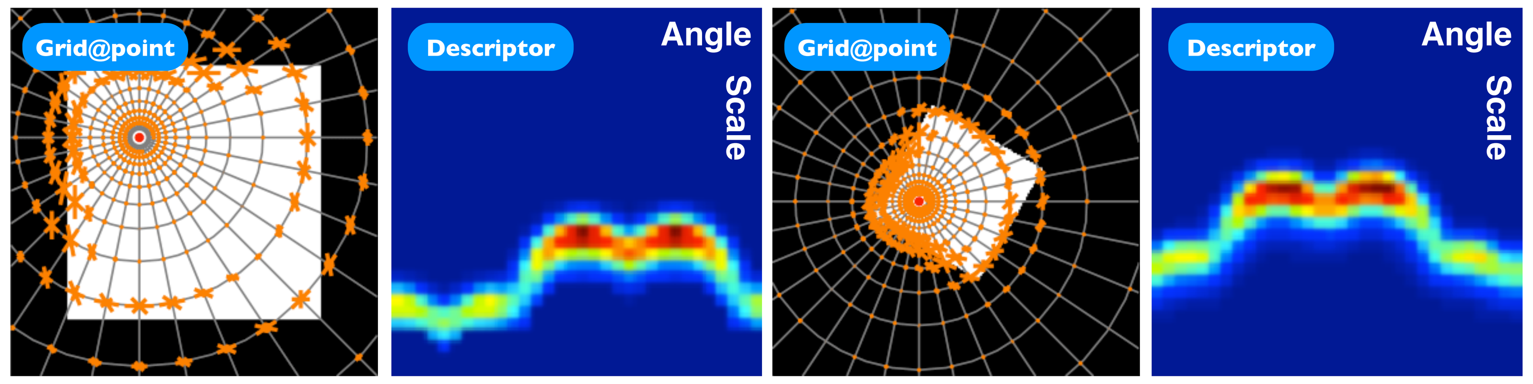


SCALE AND ROTATION INVARIANT DESCRIPTOR (SID)

Fact 1: Signal translation does not affect the signal's Fourier Transform Magnitude (shifting property).

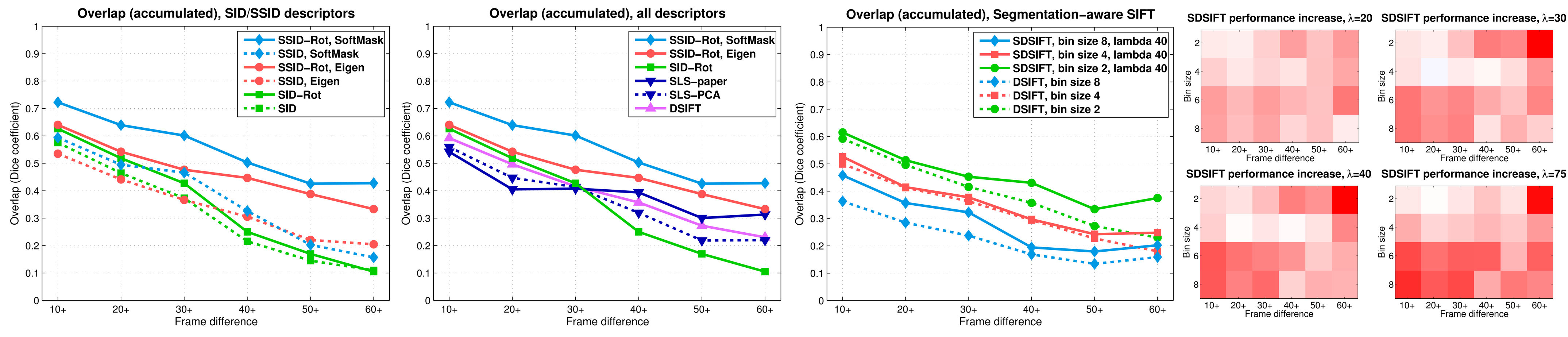
$$h[k,n] \overset{\mathcal{F}}{\leftrightarrow} H(j\omega_k, j\omega_n), h[k-c, n-d] \overset{\mathcal{F}}{\leftrightarrow} H(j\omega_k, j\omega_n)e^{-j(\omega_k c + \omega_n d)}$$

Fact 2: Log-polar sampling turns scaling and rotation to translation.



EXPERIMENT 1: LARGE DISPLACEMENT OPTICAL FLOW

MOSEG/JHU Benchmark [4]: traffic sequences with ground truth segmentation every ~10 frames.
Task: match **first and every annotated frame**. Method: SIFT-flow [5]. Metric: DICE coefficient.
Baseline: DSIFT, SID, SID-Rot, SLS [6]. **Ours:** SDSIFT, SSID and SSID-Rot with 'Eigen' or 'SoftMask'.



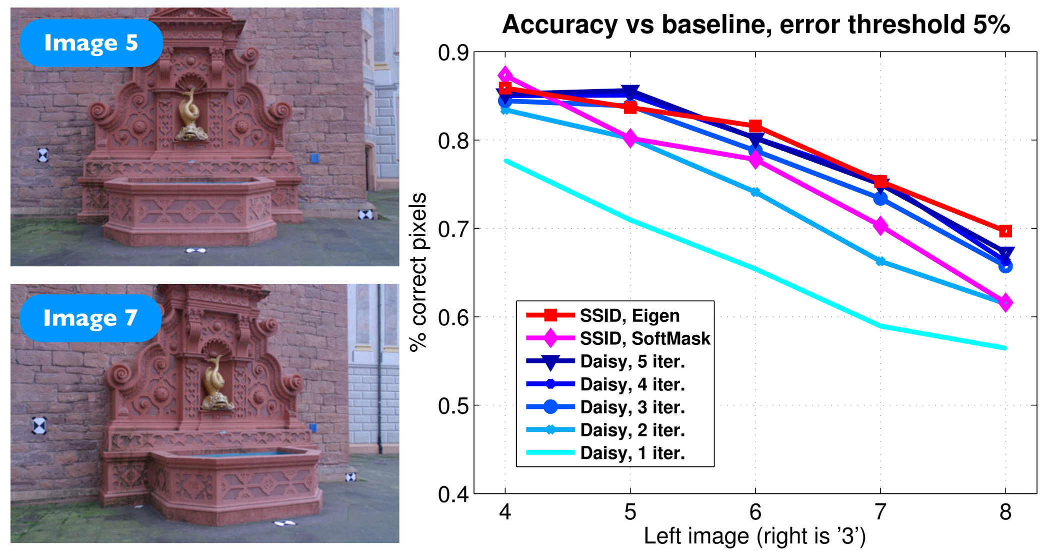
Both **SSID** and **SDSIFT** perform consistently better. **SDSIFT** has a second parameter: size.

EXP. 2: WIDE BASELINE STEREO

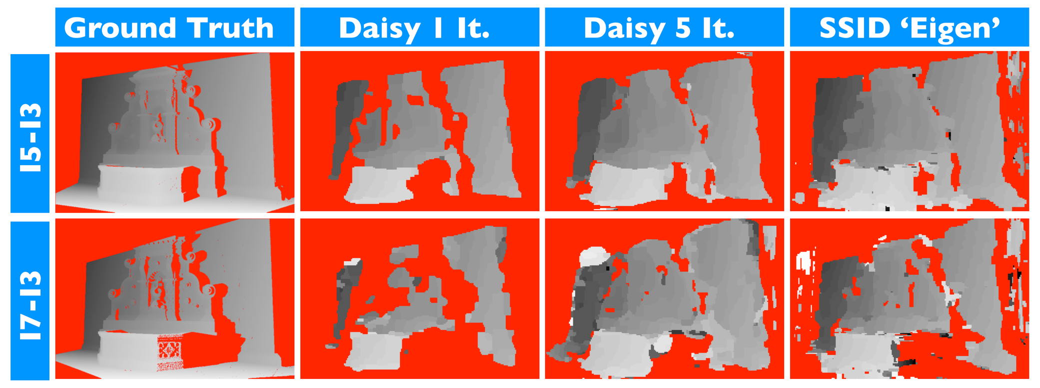
We follow the set-up of Daisy [7]:

1. Discretize 3D space into k depth layers.
2. Match subject to epipolar constraints, store best match for every depth layer.

[7]: iterative figure-ground mask estimation.
Ours: **single-shot, rotation-invariant.**



(Reference is **Image 3**: see 'Embeddings')



References

[1] I. Kokkinos, A. Yuille. Scale invariance without scale selection. CVPR 2008.
[2] M. Maire, P. Arbelaez, C. Fowlkes, J. Malik. Using contours to detect and localize junctions in natural images. CVPR 2008.
[3] M. Leordeanu, R. Sukthankar, C. Sminchisescu. Efficient closed-form solution to generalized boundary detection. ECCV 2012.
[4] T. Brox, J. Malik. Object segmentation by long term analysis of point trajectories. ECCV 2010.
[5] C. Liu, J. Yuen, A. Torralba. SIFT-flow: Dense correspondence across different scenes. PAMI 2011.
[6] T. Hassner, V. Mayzels, L. Zelnik-Manor. On SIFTS and their scales. CVPR 2012.
[7] E. Tola, V. Lepetit, P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. PAMI 2010.