SCNet: Learning Semantic Correspondence

[1]Kai Han, [4,5]Rafael S. Rezende, [2]Bumsub Ham, [1]Kwan-Yee K. Wong, [3]Minsu Cho, [4]Cordelia Schmid, [4,5]Jean Ponce

[1]HKU [2]Yonsei Univ. [3]POSTECH [4]Inria [5]ENS

ICCV 17 International Conference on Computer Vision 2017

Problem Definition and Contribution

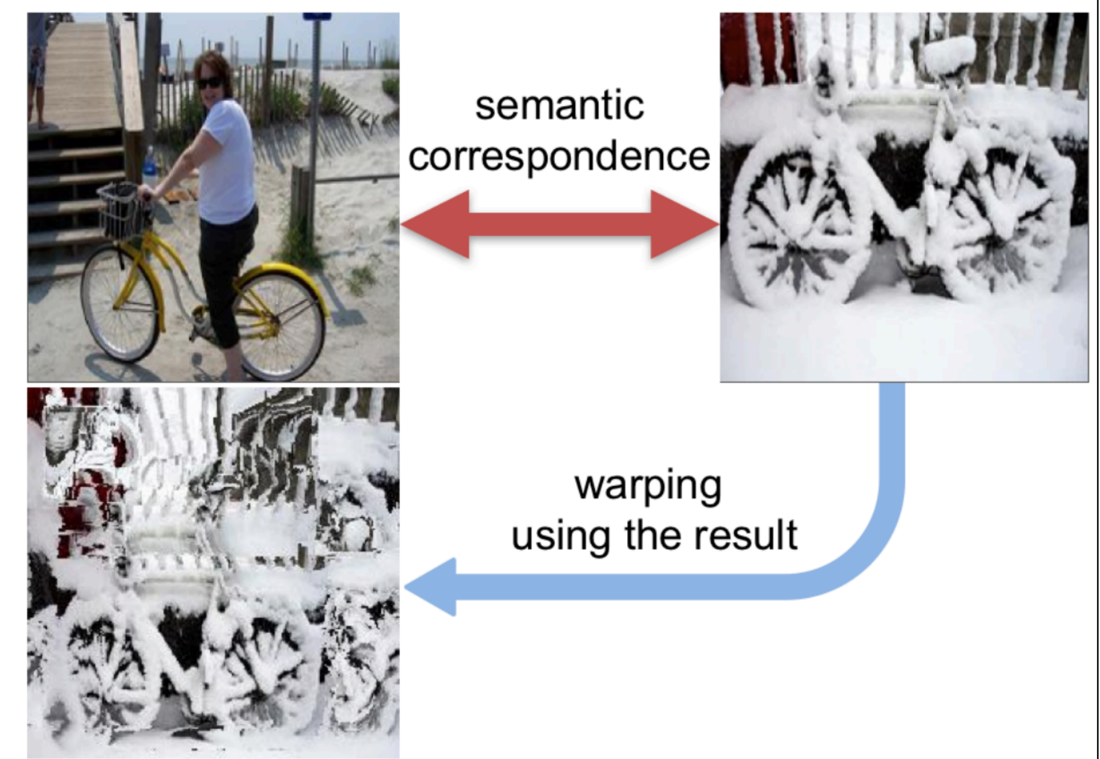
Goal: Establishing semantic correspondences between images depicting different instances of the same object or scene category.

Motivation:

- Geometric consistency constraint is a key factor in semantic matching.
- Previous approaches focus on either combining a spatial regularizer with hand-crafted features, or learning a correspondence model for appearance only.

Key contributions:

- A simple and efficient model for learning to match regions using both appearance and geometry.
- A convolutional neural network, SCNet, to learn semantic correspondence with region proposals.
- State-of-the-art results on several benchmarks, clearly demonstrating the advantage of learning both appearance and geometric terms.



Problem Formulation

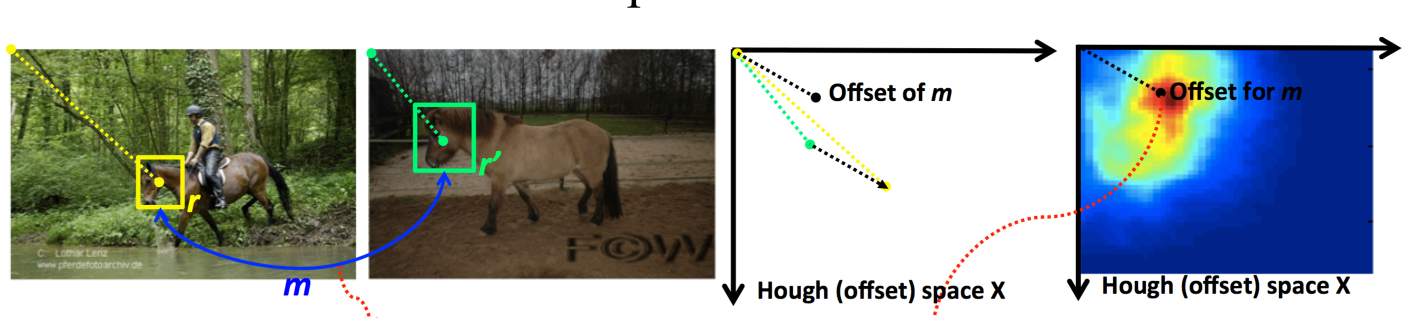
Probabilistic Hough matching (PHM) [1, 2]:

Region $\boldsymbol{r} = (\boldsymbol{f}, \boldsymbol{l})$: feature $\boldsymbol{f}$ and location $\boldsymbol{l}$

Data $\boldsymbol{D} = (\boldsymbol{R}, \boldsymbol{R'})$: two sets of regions $\boldsymbol{R}$ and $\boldsymbol{R'}$

Match $\boldsymbol{m} = (\boldsymbol{r}, \boldsymbol{r'})$: a pair of regions in $\boldsymbol{R} \times \boldsymbol{R'}$

Offset of $\boldsymbol{m}$ as $\boldsymbol{x} = \boldsymbol{l} - \boldsymbol{l'}$: displacement between $\boldsymbol{r}$ and $\boldsymbol{r'}$



$$P(m|D) \approx \underbrace{P_a(m)}_{\text{Appearance}} \underbrace{\sum_x P_g(m|x) \sum_{m' \in D} P_a(m') P_g(m'|x)}_{\text{Geometry}}$$

In our learning framework, we consider similarity rather than probabilities:

$$\begin{aligned} z(m,w) &= f(m,w) \sum_x g(m,x) \sum_{m' \in D} f(m',w) g(m',x) \\ &= f(m,w) \sum_{m' \in D} [\sum_x g(m,x) g(m',x)] f(m',w) \end{aligned}$$

$$z(m,w) = f(m,w) \sum_{m'} K_{mm'} f(m',w),$$

$$\text{where } K_{mm'} = \sum_x g(m,x) g(m',x)$$

x runs over a grid of predefined offset values, and $h(m)$ assigns match m to the nearest offset point.

$$K_{mm'} = \begin{cases} 1, & \text{if } h(m) = h(m') \\ 0, & \text{otherwise.} \end{cases}$$

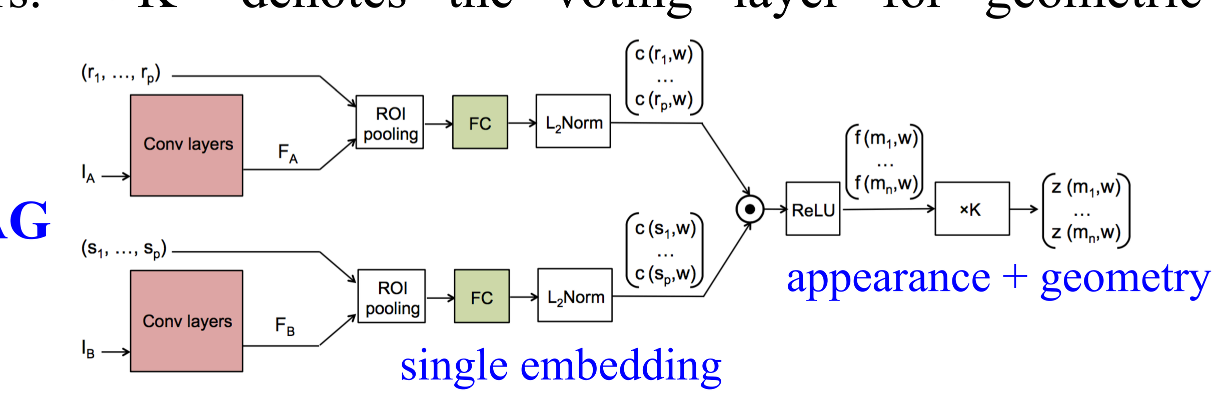
We can learn our similarity function by minimizing w.r.t the network parameters w:

$$E(w) = \sum_{m=1}^{n} l[y_m, z(m,w)] + \lambda \Omega(w)$$

SCNet Architectures

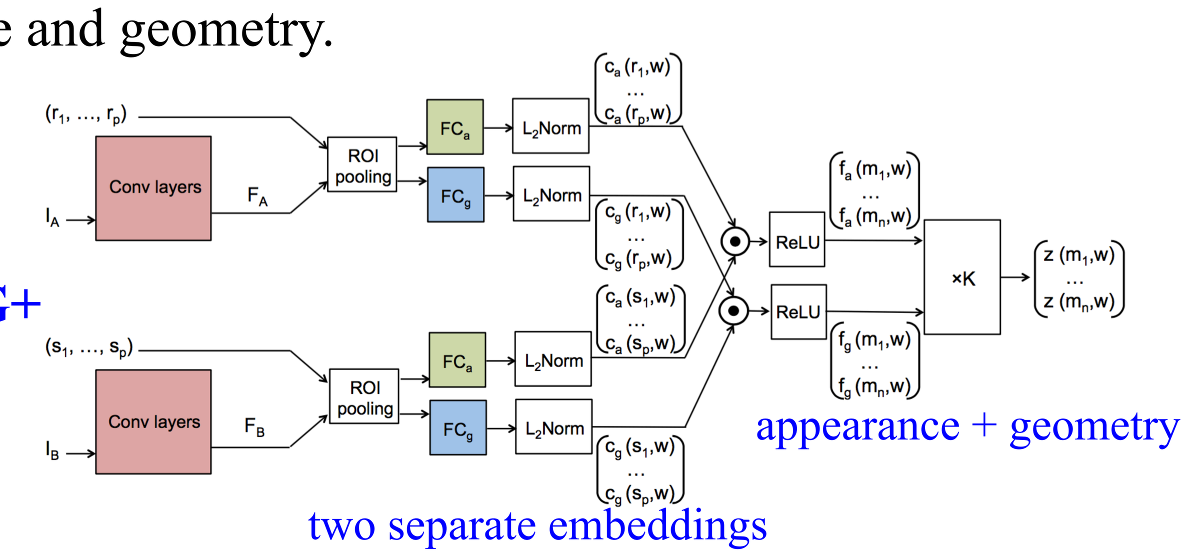
Three variants are proposed: SCNet-AG, SCNet-AG+, and SCNet-A. Colored boxes represent layers with learning parameters and the boxes with the same color share the same parameters. "×K" denotes the voting layer for geometric scoring.

SCNet-AG



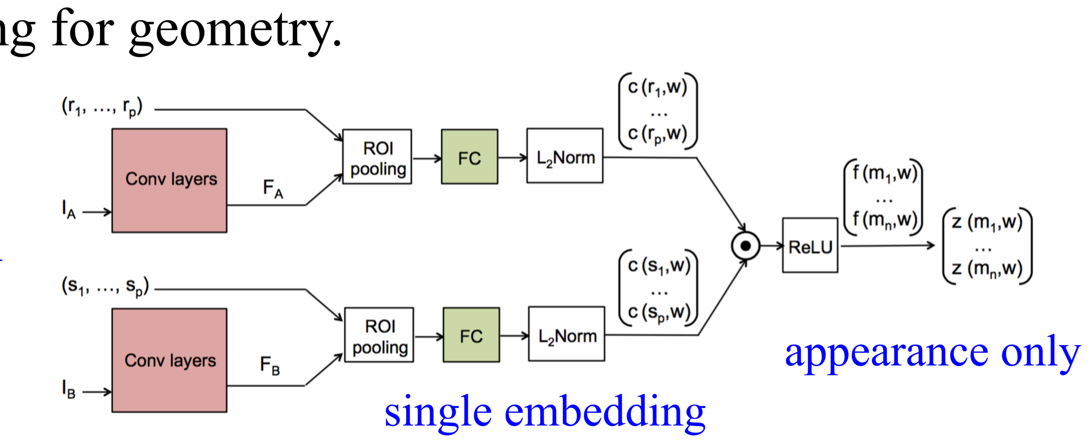
The basic architecture. It learns a single embedding for both appearance and geometry.

SCNet-AG+



An extended variant. It learns an additional and separate embedding for geometry.

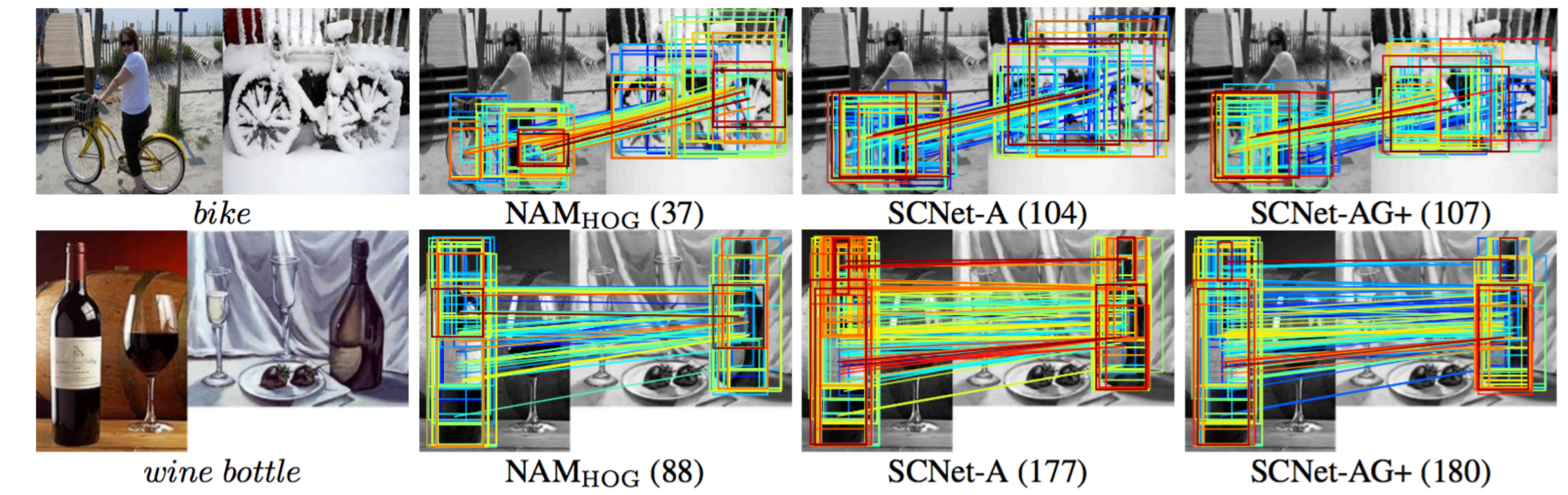
SCNet-A



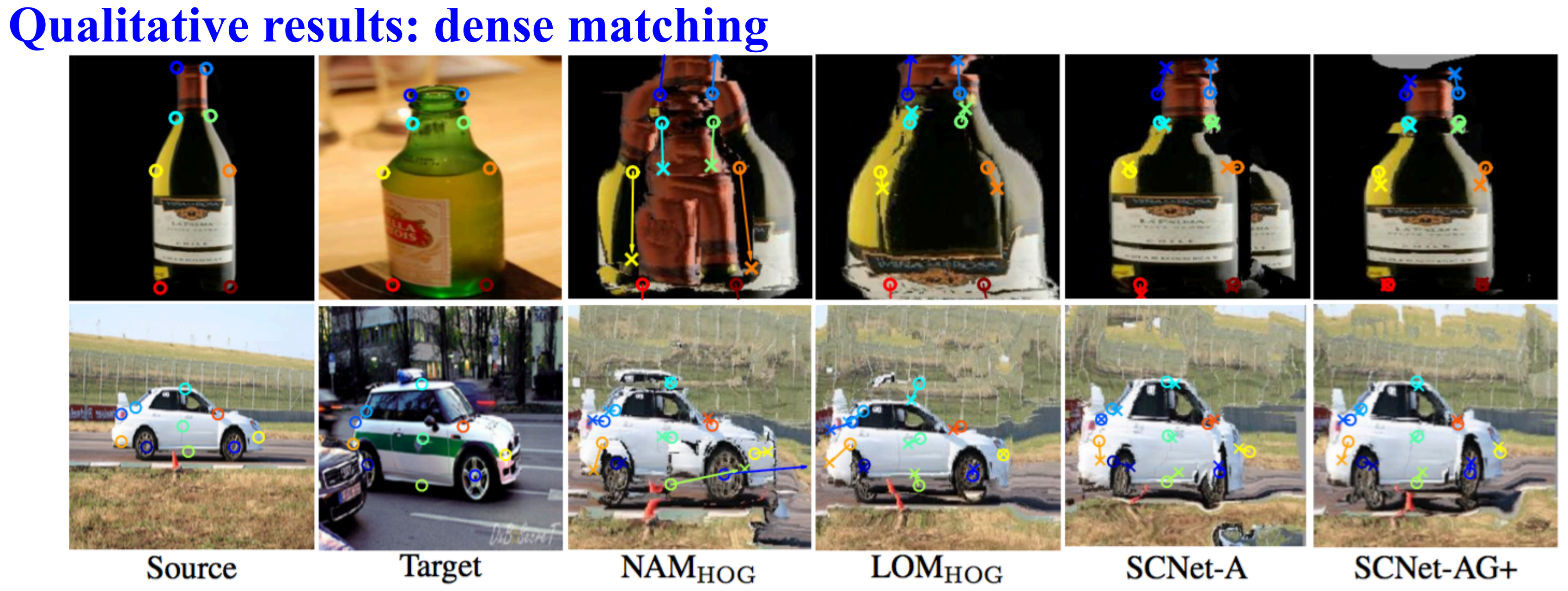
A simplified variant. It learns appearance information only by making the voting layer an identity function.

Experiments

Qualitative results: region matching



Qualitative results: dense matching



Quantitative results

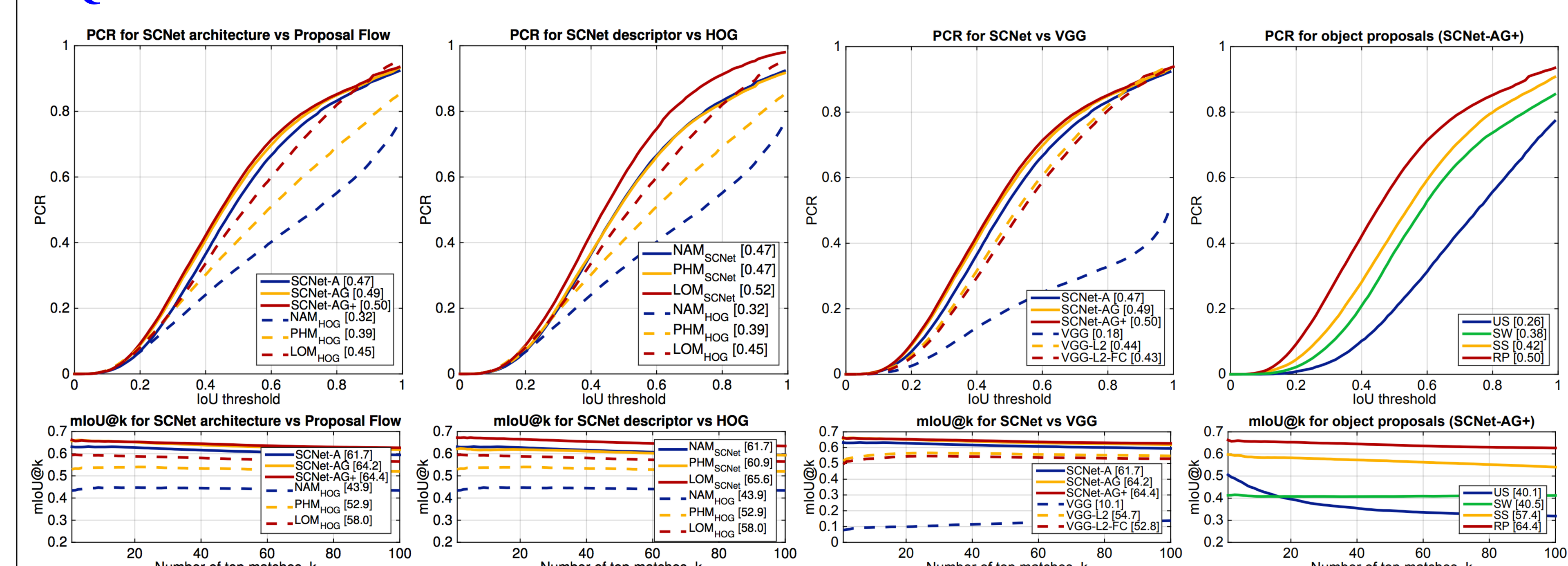


Table 1: Per-class PCK on PF-PASCAL at $\tau = 0.1$. For all methods using object proposals, we use 1000 RP proposals.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	d.table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mean
NAM_{HOG}	72.9	73.6	31.5	52.2	37.9	71.7	71.6	34.7	26.7	48.7	28.3	34.0	50.5	61.9	26.7	51.7	66.9	48.2	47.8	59.0	52.5
PHM_{HOG}	78.3	76.8	48.5	46.7	45.9	72.5	72.1	47.9	49.0	84.0	37.2	46.5	51.3	72.7	38.4	53.6	67.2	50.9	60.0	63.4	60.3
LOM_{HOG}	73.3	74.4	54.4	50.9	49.6	73.8	72.9	63.6	46.1	79.8	42.5	48.0	68.3	66.3	42.1	62.1	65.2	57.1	64.4	58.0	62.5
UCN	64.8	58.7	42.8	59.6	47.0	42.2	61.0	45.6	49.9	52.0	48.5	49.5	53.2	72.7	53.0	41.4	**83.3**	49.0	**73.0**	66.0	55.6
SCNet-A	67.6	72.9	69.3	59.7	**74.5**	72.7	73.2	59.5	51.4	78.2	39.4	50.1	67.0	62.1	**69.3**	**68.5**	78.2	63.3	57.7	59.8	66.3
SCNet-AG	83.9	81.4	**70.6**	62.5	60.6	81.3	81.2	59.5	53.1	81.2	**62.0**	58.7	65.5	73.3	51.2	58.3	60.0	69.3	61.5	**80.0**	69.7
SCNet-AG+	**85.5**	**84.4**	66.3	**70.8**	57.4	**82.7**	**82.3**	**71.6**	**54.3**	**95.8**	55.2	**59.5**	**68.6**	**75.0**	56.3	60.4	60.0	**73.7**	66.5	76.7	**72.2**

Table 2: Fixed-threshold PCK on PF-WILLOW.

Method	PCK@0.05	PCK@0.1	PCK@0.15
SIFT Flow	0.247	0.380	0.504
DAISY w/SF	0.324	0.456	0.555
DeepC w/SF	0.212	0.364	0.518
LIFT w/SF	0.224	0.346	0.489
VGG w/SF	0.224	0.388	0.555
FCSS w/SF	0.354	0.532	0.681
FCSS w/PF	0.295	0.584	0.715
LOM_{HOG}	0.284	0.568	0.682
UCN	0.291	0.417	0.513
SCNet-A	0.390	**0.725**	**0.873**
SCNet-AG	**0.394**	0.721	0.871
SCNet-AG+	0.386	0.704	0.853

Table 3: Results on Caltech-101.

Methods	LT-ACC	IoU	LOC-ERR
NAM_{HOG}	0.70	0.44	0.39
PHM_{HOG}	0.75	0.48	0.31
LOM_{HOG}	0.78	0.50	0.26
DeepFlow	0.74	0.40	0.34
SIFT Flow	0.75	0.48	0.32
DSP	0.77	0.47	0.35
FCSS w/SF	0.80	0.50	**0.21**
FCSS w/PF	**0.83**	**0.52**	0.22
SCNet-A	0.78	0.50	0.28
SCNet-AG	0.78	0.50	0.27
SCNet-AG+	0.79	0.51	0.25

Table 4: Results on PASCAL Parts.

Methods	IoU	PCK
NAM_{HOG}	0.35	0.13
PHM_{HOG}	0.39	0.17
LOM_{HOG}	0.41	0.17
Congealing	0.38	0.11
RASL	0.39	0.16
CollectionFlow	0.38	0.12
DSP	0.39	0.17
FCSS w/SF	0.44	0.28
FCSS w/PF	0.46	**0.29**
SCNet-A	0.47	0.17
SCNet-AG	0.47	0.17
SCNet-AG+	**0.48**	0.18

Project webpage: http://www.di.ens.fr/willow/research/scnet/

[1] M. Cho, S. Kwak, C. Schmid, J. Ponce, Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals, CVPR 2015

[2] B. Ham, M. Cho, C. Schmid, J. Ponce, Proposal Flow, CVPR 2016