Part I
Unsupervised Feature Learning with Convolutional Neural Networks

Thomas Brox
Computer Vision Group
University of Freiburg, Germany

Research funded by ERC Starting Grant VideoLearn and Deutsche Telekom Stiftung

Status quo: CNNs generate great features

Team name	Error (5 guesses)	Description
SuperVision	0.15315	Using extra training data from ImageNet Fall 2011 release
SuperVision	0.16422	Using only supplied training data
ISI	0.26172	Weighted sum of scores from each classifier with SIFT+FV, LBP+FV, GIST+FV, and CSIFT+FV, respectively.

ILSVRC 2012 classification
Krizhevsky et al. 2012

VOC 2007 test	aero	bike	bird		tv	mAP
R-CNN FT fc_7 BB	**68.1**	**72.8**	**56.8**		**64.8**	**58.5**
DPM v5	33.2	60.3	10.2	...	43.5	33.7
DPM ST	23.8	58.2	10.5		44.9	29.1
DPM HSC	32.2	58.3	11.5		45.2	34.3

PASCAL VOC object detection
Girshick et al. 2014

Do we need these massive amounts
of class labels to learn generic features?

Unsupervised feature learning

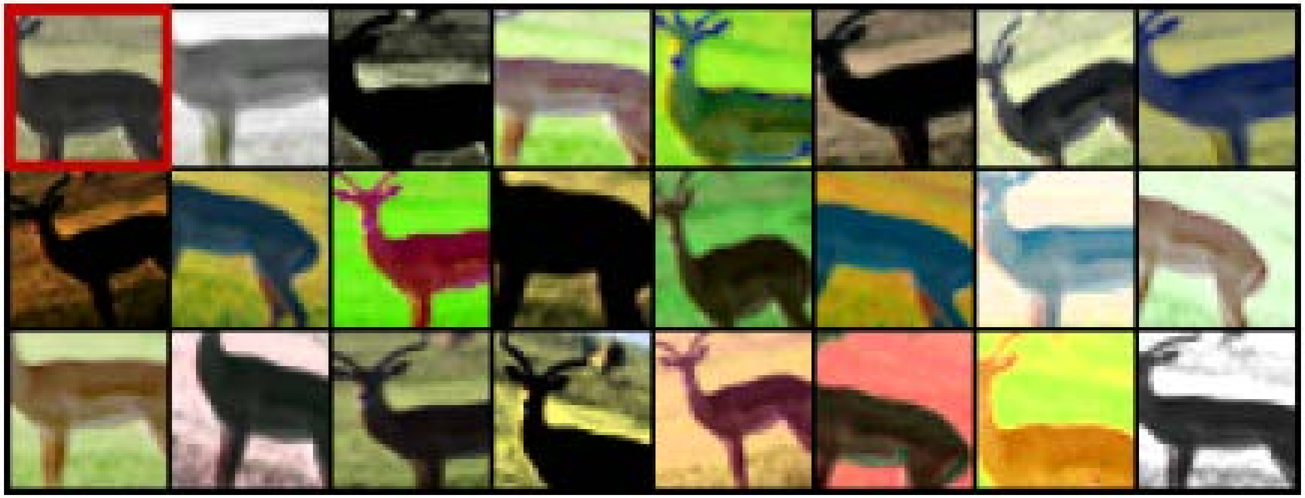
- Dominant concept: reconstruction error + regularization

- Existing frameworks:
  - Autoencoders (dimensionality reduction)
    (Hinton 1989, Vincent et al. 2008,…)
  - Sparse coding (sparsity prior)
    (Olshausen-Field 1996, Mairal et al. 2009, Bo et al. 2012,…)
  - Slowness prior
    (Wiscott-Sejnowski 2002, Zou et al. 2012,…)
  - Deep belief networks (prior in contrastive divergence)
    (Ranzato et al. 2007, Lee et al. 2009,…)

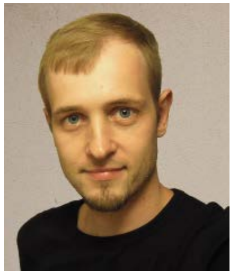
- Reconstruction error models the input distribution
  → dubious objective

Exemplar CNN: discriminative objective

- Train CNN to discriminate **surrogate classes**



Alexey Dosovitskiy



Jost Tobias Springenberg

- Take data augmentation to the extreme
  (translation, rotation, scaling, color, contrast, brightness)

- Transformations define invariance properties of the features to be learned

Acknowledgements to caffe.berkeleyvision.org

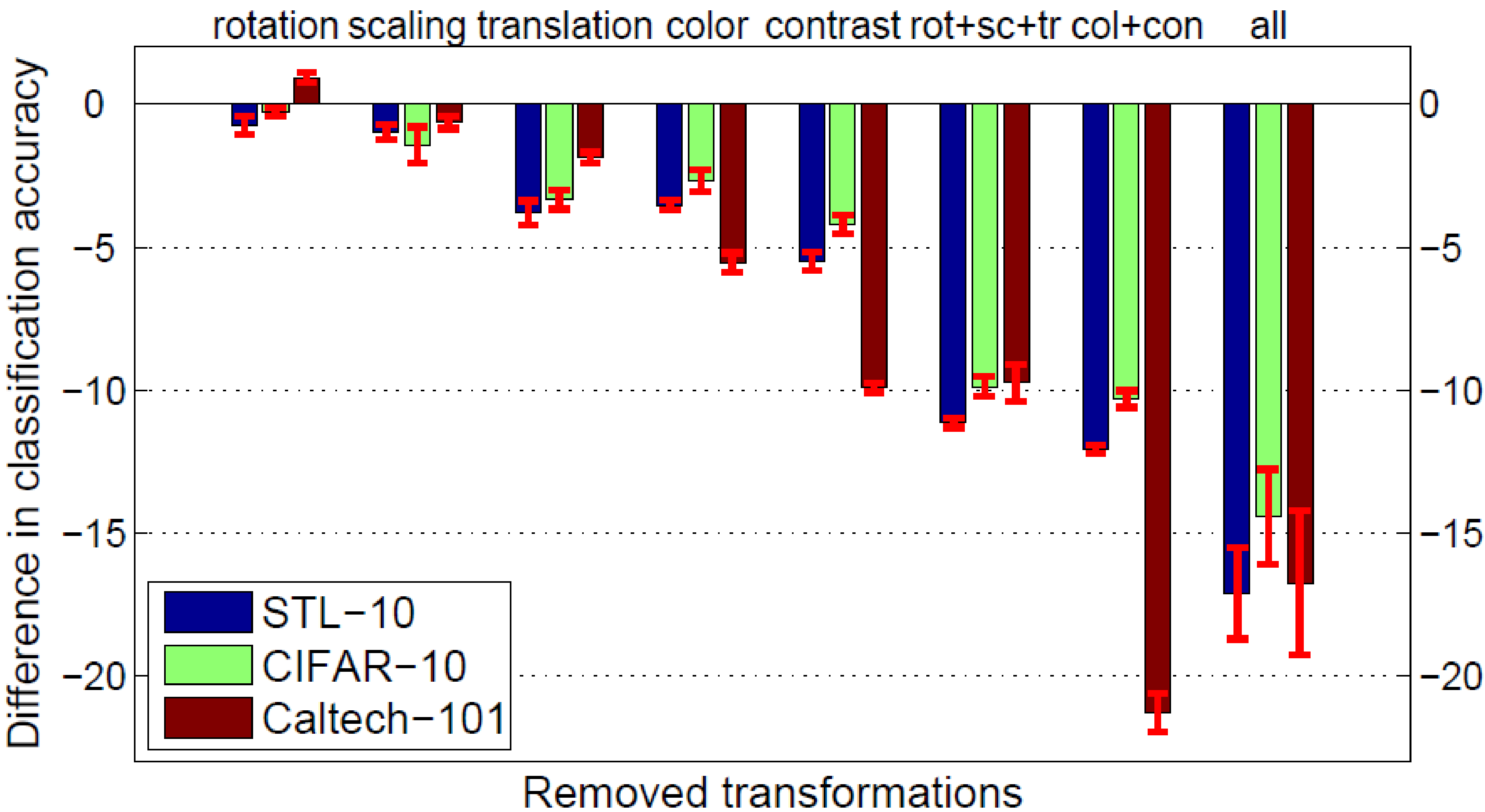
Application to classification

- Pooled responses from each layer used as features
- Training of linear SVM

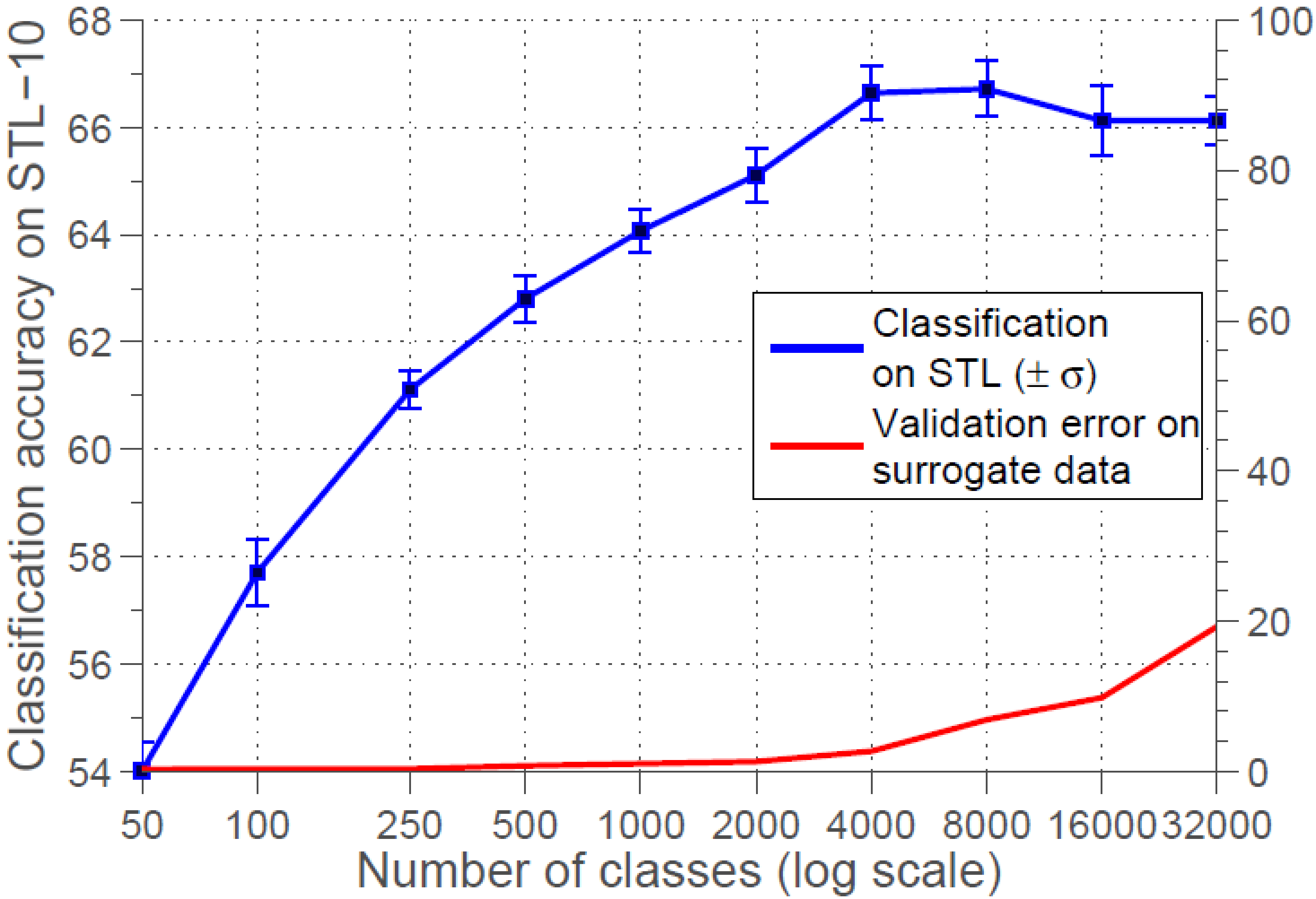
	STL-10	CIFAR-10	Caltech-101
Convolutional K-means network	60.1	70.7	-
View-invariant K-means	63.7	72.6	-
Multi-way local pooling	-	-	77.3
Slowness on video	61.0	-	74.6
Hierarchical Matching Pursuit (HMP)	64.5	-	-
Multipath HMP	-	-	82.5
Exemplar CNN	**72.8**	**75.3**	**85.5**

Outperforms all previous unsupervised feature learning approaches

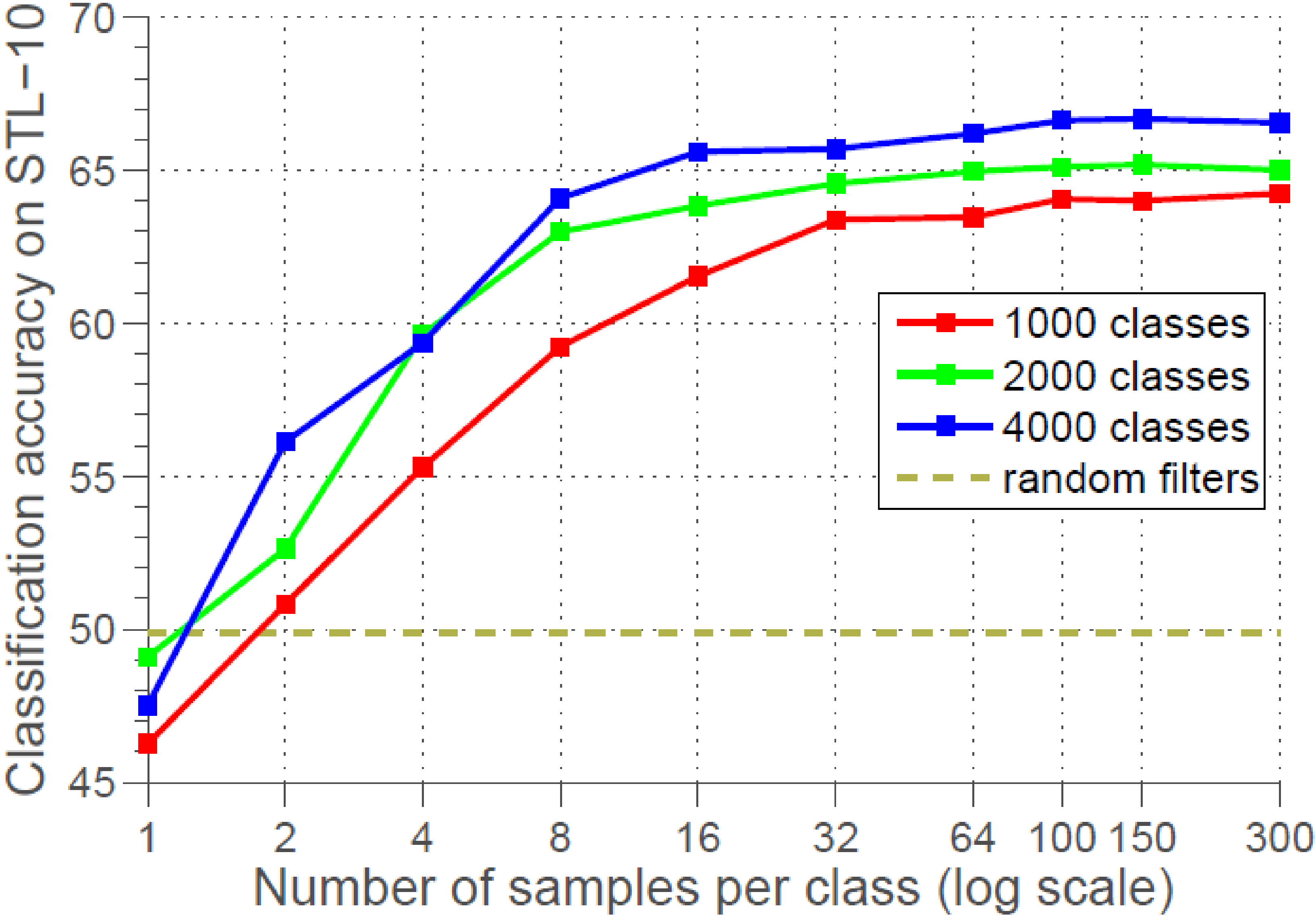
Which transformations are most relevant?



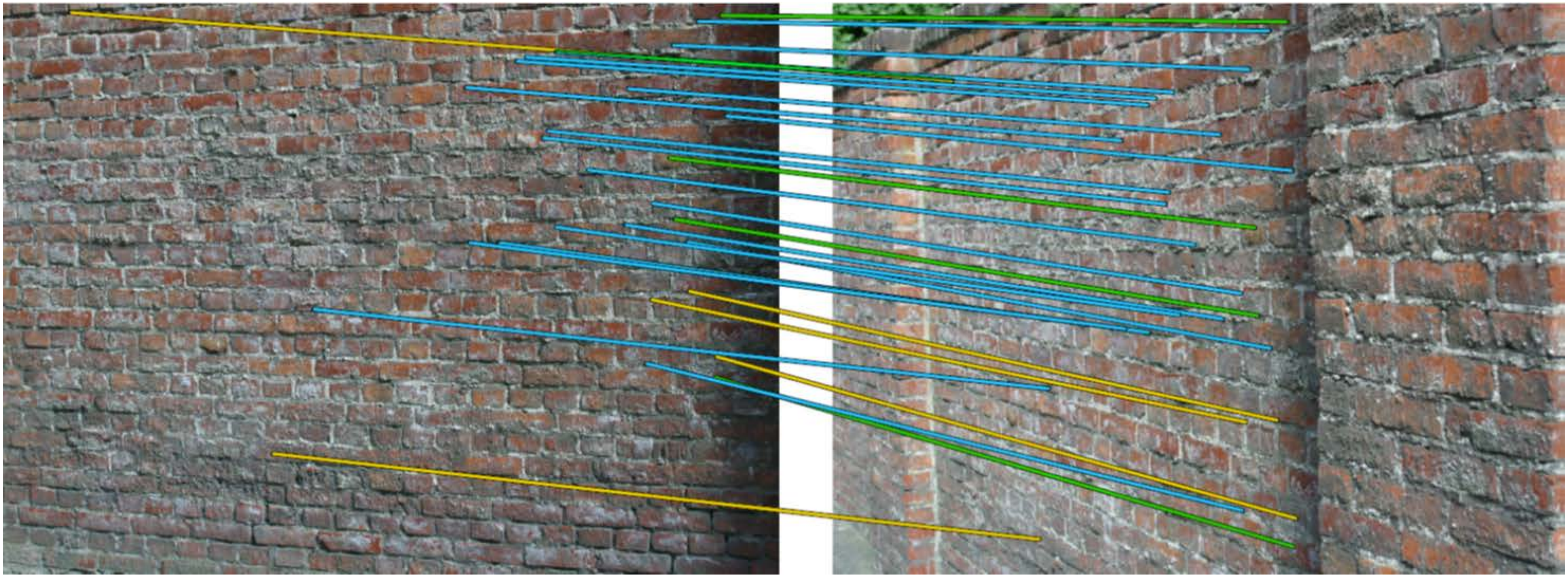
How many surrogate classes?



How many samples per class?

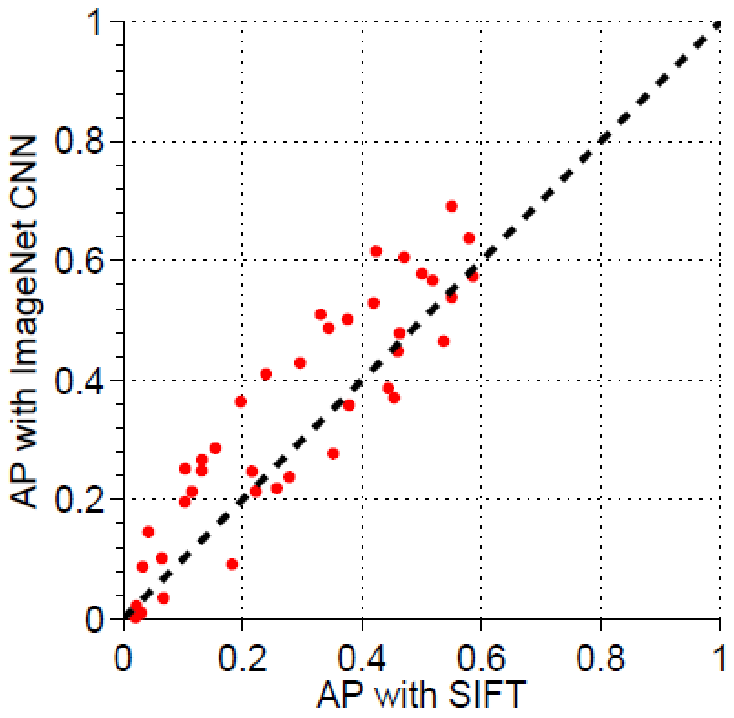


Application to descriptor matching

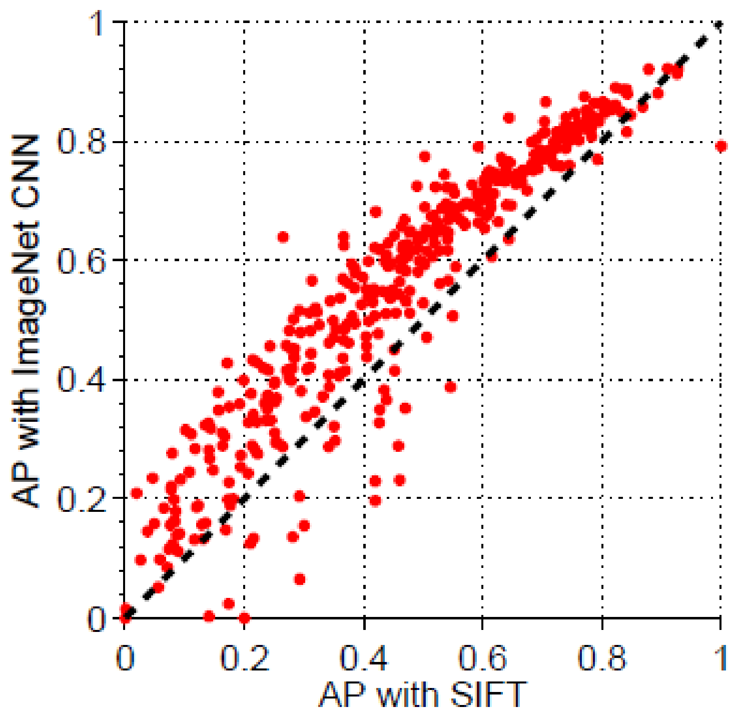


Descriptor matching between two images

CNNs won’t work for descriptor matching, right?



Mikolajczyk dataset



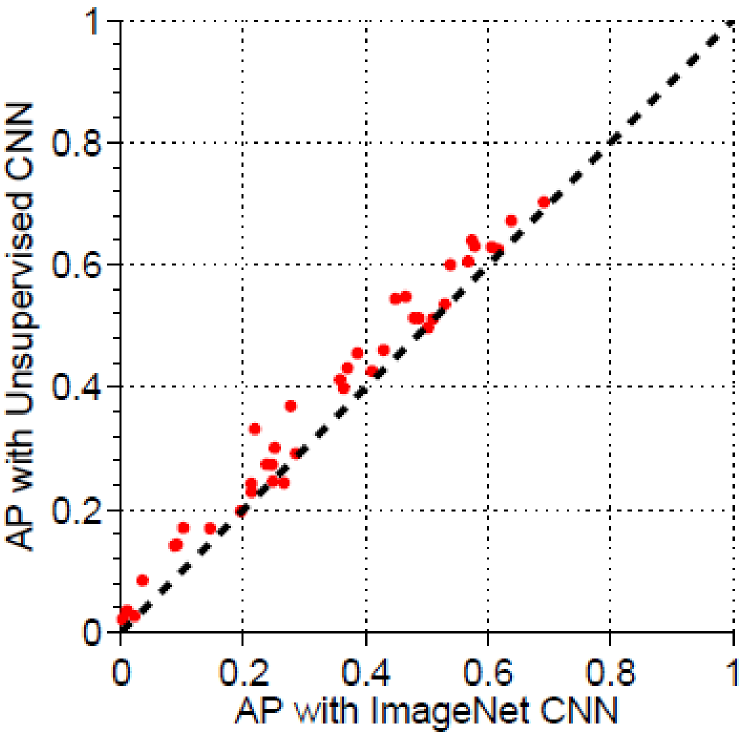
New larger dataset

Philipp Fischer

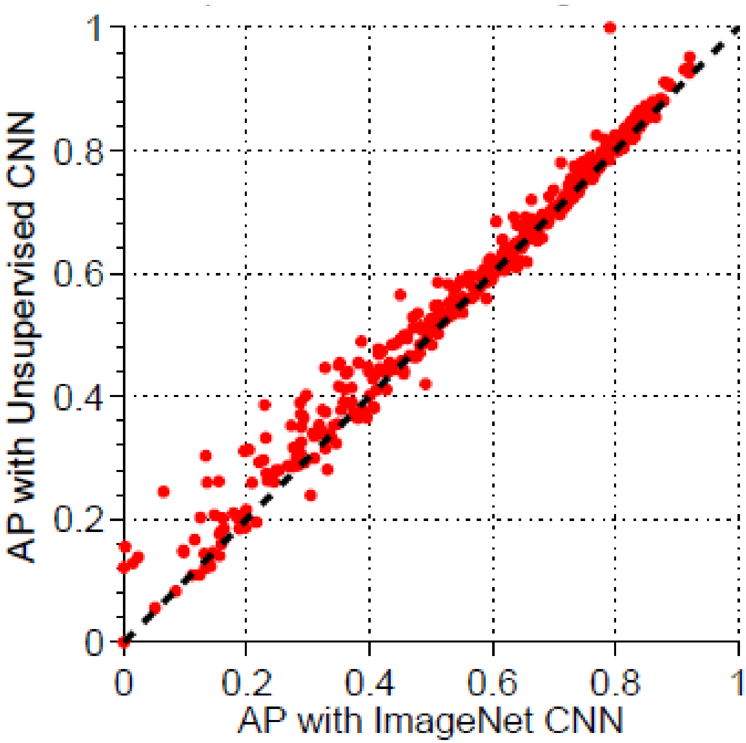
Alexey Dosovitskiy

Descriptors from a CNN outperform SIFT

Supervised versus unsupervised CNN



Mikolajczyk dataset



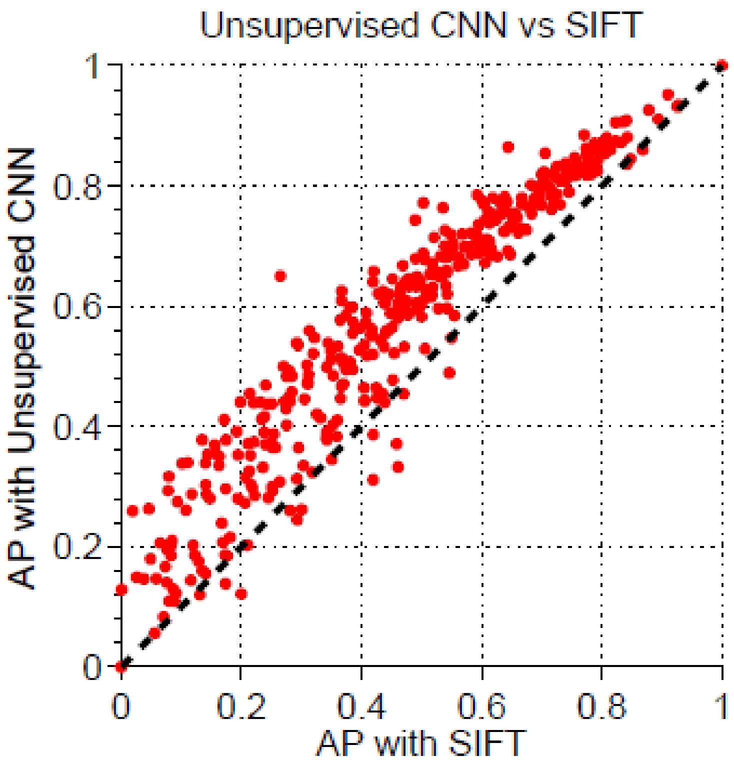
New larger dataset

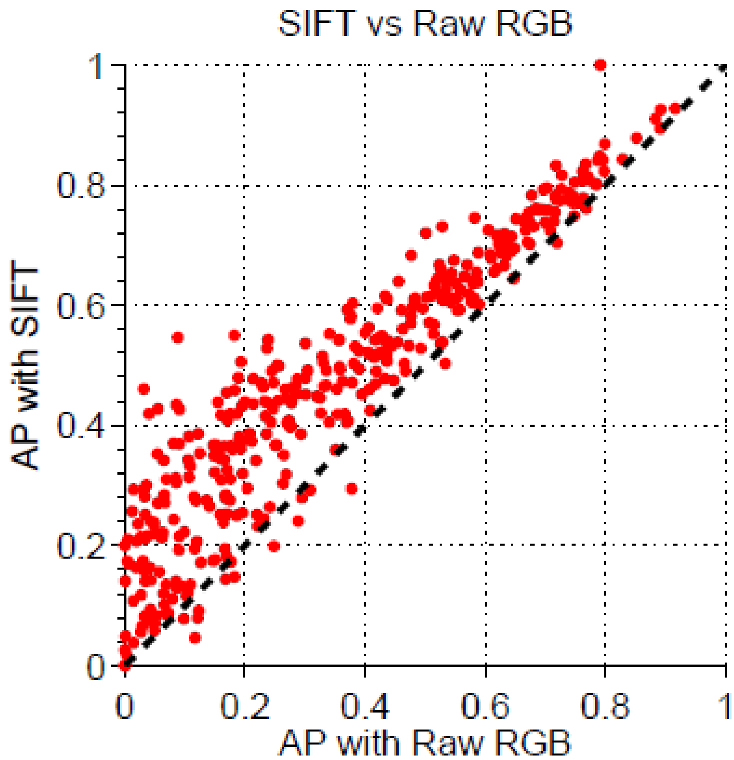
Philipp Fischer

Alexey Dosovitskiy

Unsupervised feature learning advantageous for descriptor matching

Relevance of improvement



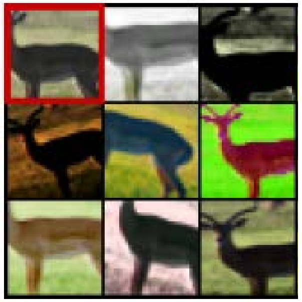


Philipp Fischer

Alexey Dosovitskiy

Improvement of Examplar CNN over SIFT
is as big as SIFT over color patches

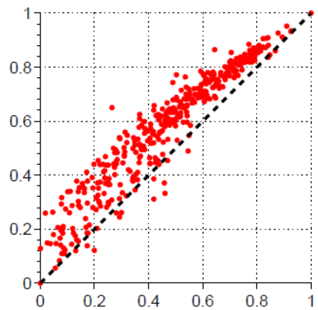
Summary of part I



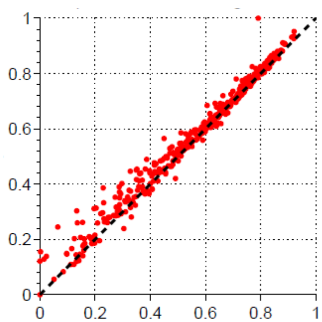
Exemplar CNN: Unsupervised feature learning by discriminating surrogate classes

	STL-10	CIFAR-10
Convolutional K-means network	60.1	70.7
View-invariant K-means	63.7	72.6
Multi-way local pooling	-	-
Slowness on video	61.0	-
Hierarchical Matching Pursuit (HMP)	64.5	-
Multipath HMP	-	-
Surrogate Class CNN	**72.8**	**75.3**

Outperforms previous unsupervised methods on classification



CNNs outperform SIFT even on descriptor matching



Unsupervised training advantageous for descriptor matching

Part II
Benchmarking Video Segmentation

Thomas Brox
Computer Vision Group
University of Freiburg, Germany

Contains joint work with
Fabio Galasso, Bernt Schiele (MPI Saarbrücken)

Research funded by DFG and ERC

Motion segmentation

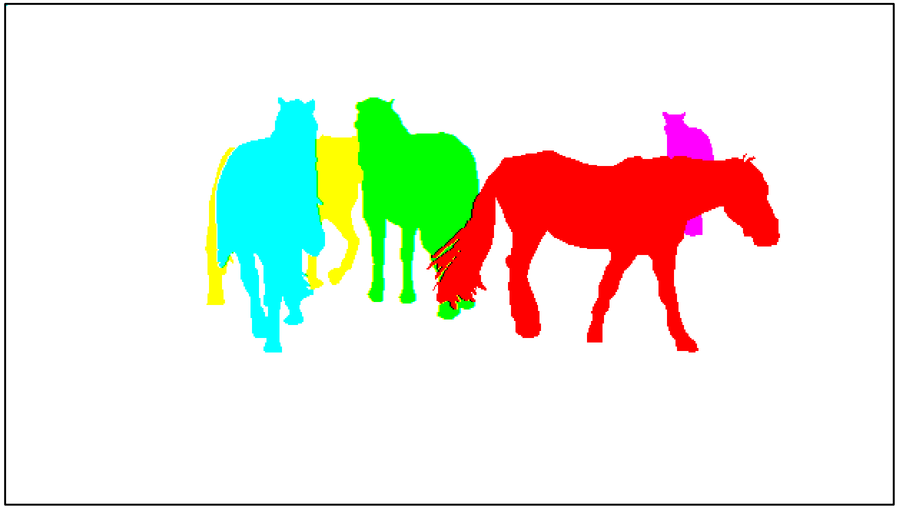
Brox-Malik
ECCV 2010

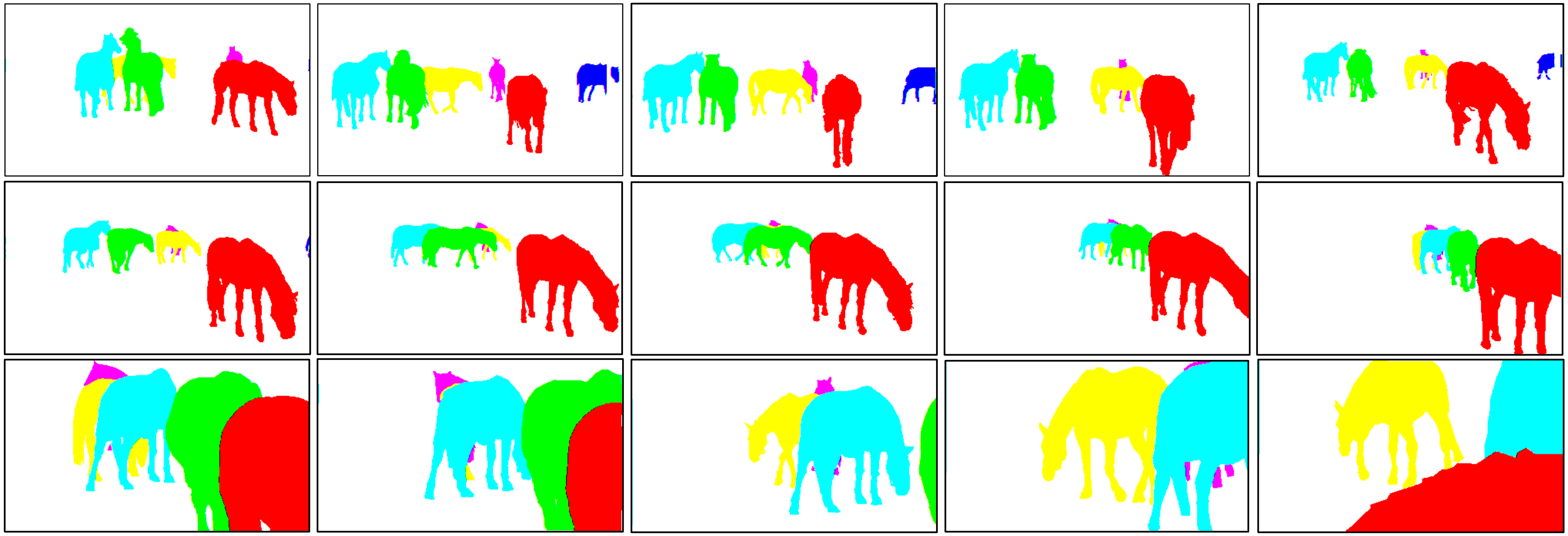
Ochs et al.
PAMI 2014

Benchmarking motion segmentation

Freiburg-Berkeley Motion Segmentation Dataset (FBMS-59)
59 sequences split into a training and a test set

Pixel-accurate ground truth



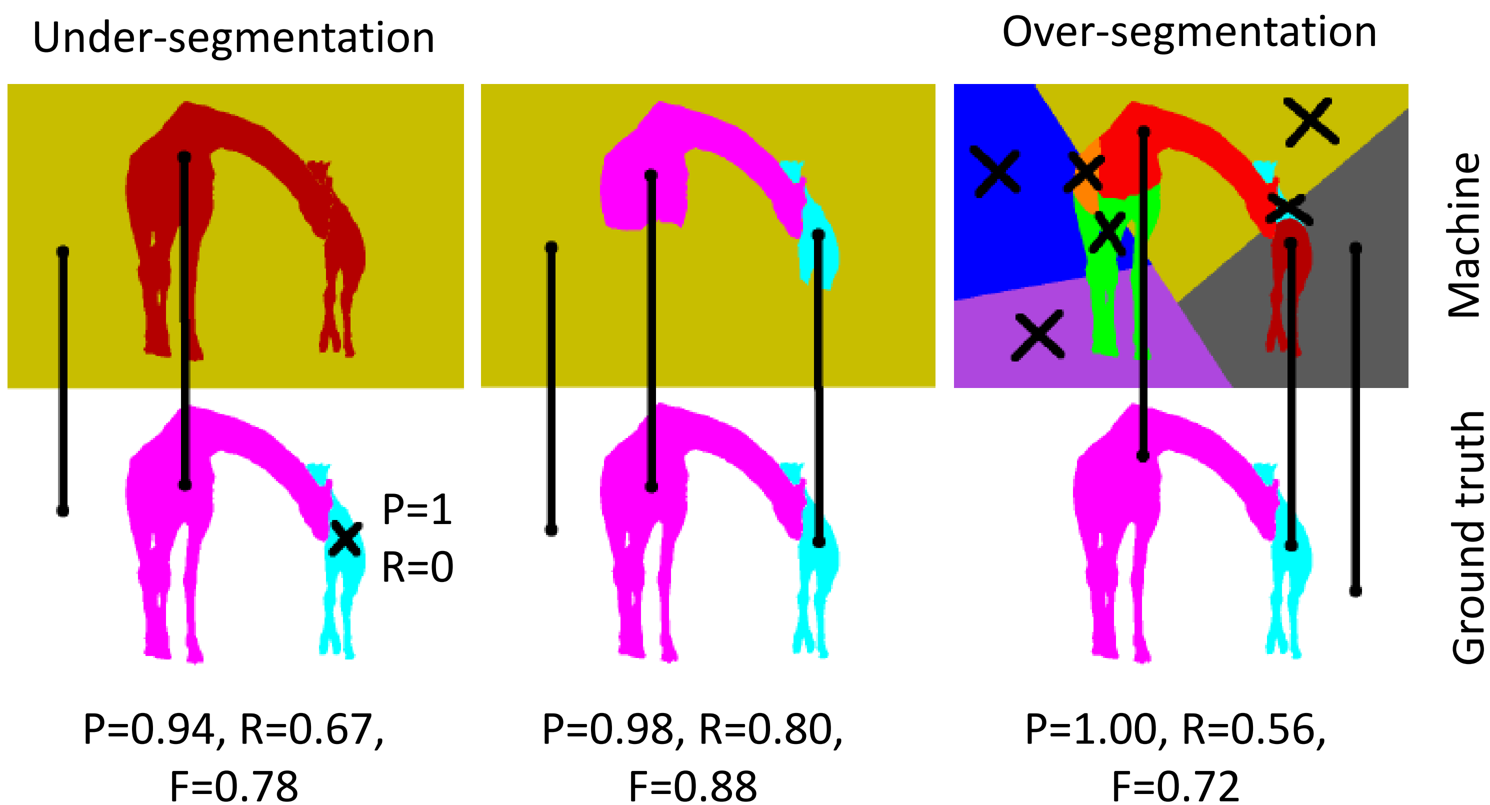


...

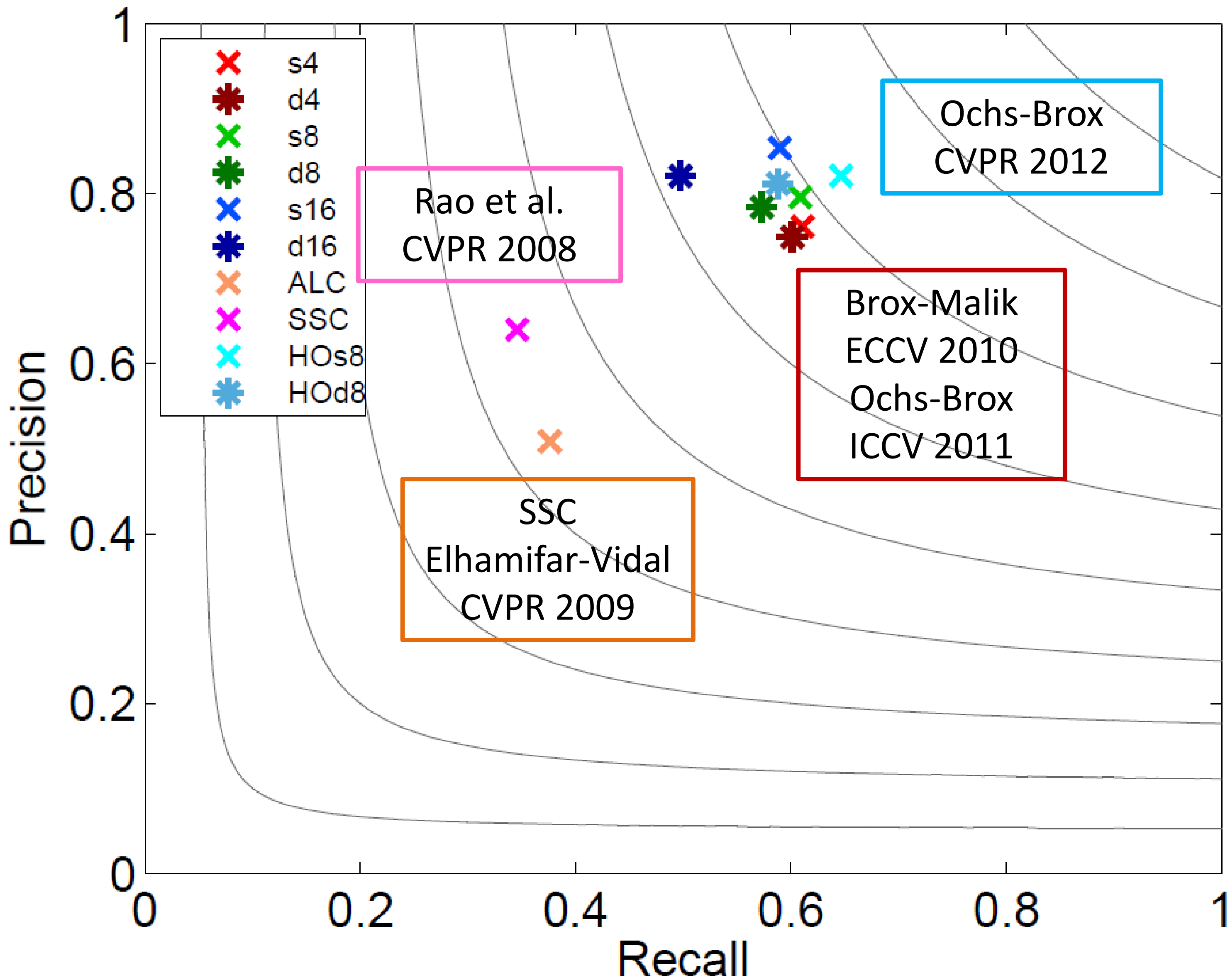
Ground truth mostly every 20 frames

Region c_j to ground truth g_i assignment with Hungarian method

$$P = \frac{1}{n} \sum_{i=1}^{n} \frac{c_j \cap g_i}{c_j} \qquad R = \frac{1}{n} \sum_{i=1}^{n} \frac{c_j \cap g_i}{g_i} \qquad F = \frac{2PR}{P+R}$$



Results on the test set



Ochs et al.
PAMI 2014

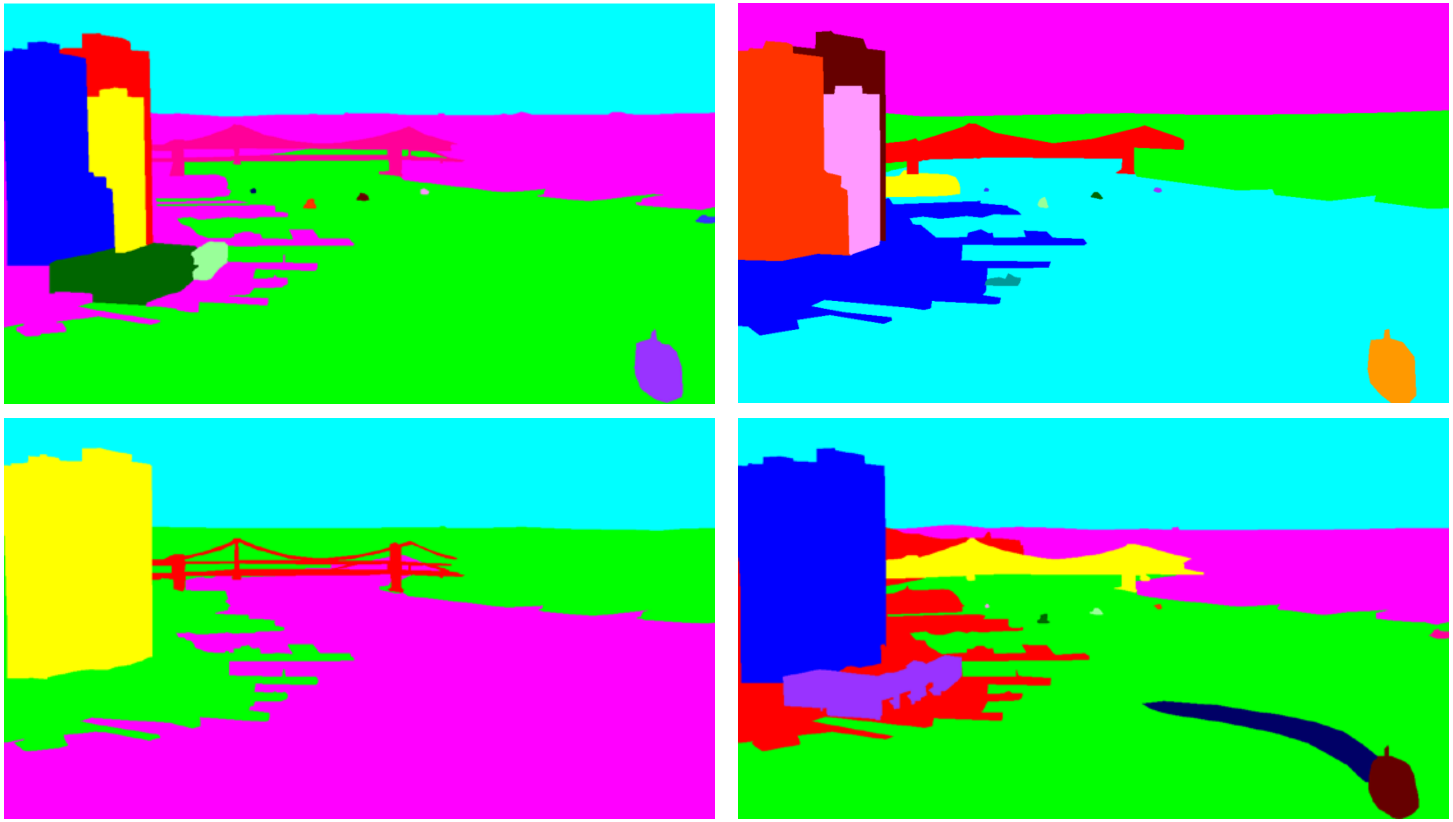
Benchmarking general video segmentation

VSB100

VSB-100: Benchmark based on Berkeley Video Segmentation Dataset
100 HD videos (40 training, 60 test)

Four human annotations per video

VSB100



Fabio Galasso

Naveen S. Nagaraja

Bernt Schiele

Galasso et al. ICCV 13

Metric for supervoxels

VSB100

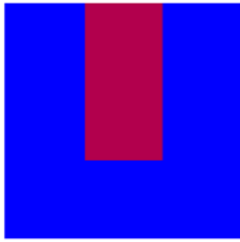
Average over all human annotations

For each region find ground truth with max overlap

Normalize by size of largest ground truth region (single region yields P=0)

$$P = \frac{\frac{1}{H}\sum_{i=1}^{H}\left(\left(\sum_{c\in\mathcal{C}} \max_{g\in\mathcal{G}_i} |c\cap g|\right) \boxed{- \max_{g\in\mathcal{G}_i}|g|}\right)}{\sum_{c\in\mathcal{C}}|c| \boxed{-\frac{1}{H}\sum_{i=1}^{H} \max_{g\in\mathcal{G}_i}|g|}}$$

Evaluated pixels in the video minus the largest ground truth region



GT

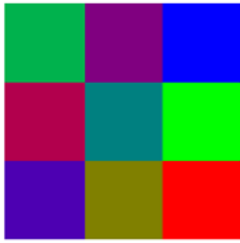
Average over all human annotations

For each ground truth find region with max overlap

$$R = \frac{\sum_{i=1}^{H}\left(\sum_{g\in\mathcal{G}_i}\max_{c\in\mathcal{C}}|c\cap g| \boxed{-1}\right)}{\sum_{i=1}^{H}\left(\sum_{g\in\mathcal{G}_i}|g|\boxed{-|\mathcal{G}_i|}\right)}$$

Size of all ground truth regions minus size of the largest ground truth region

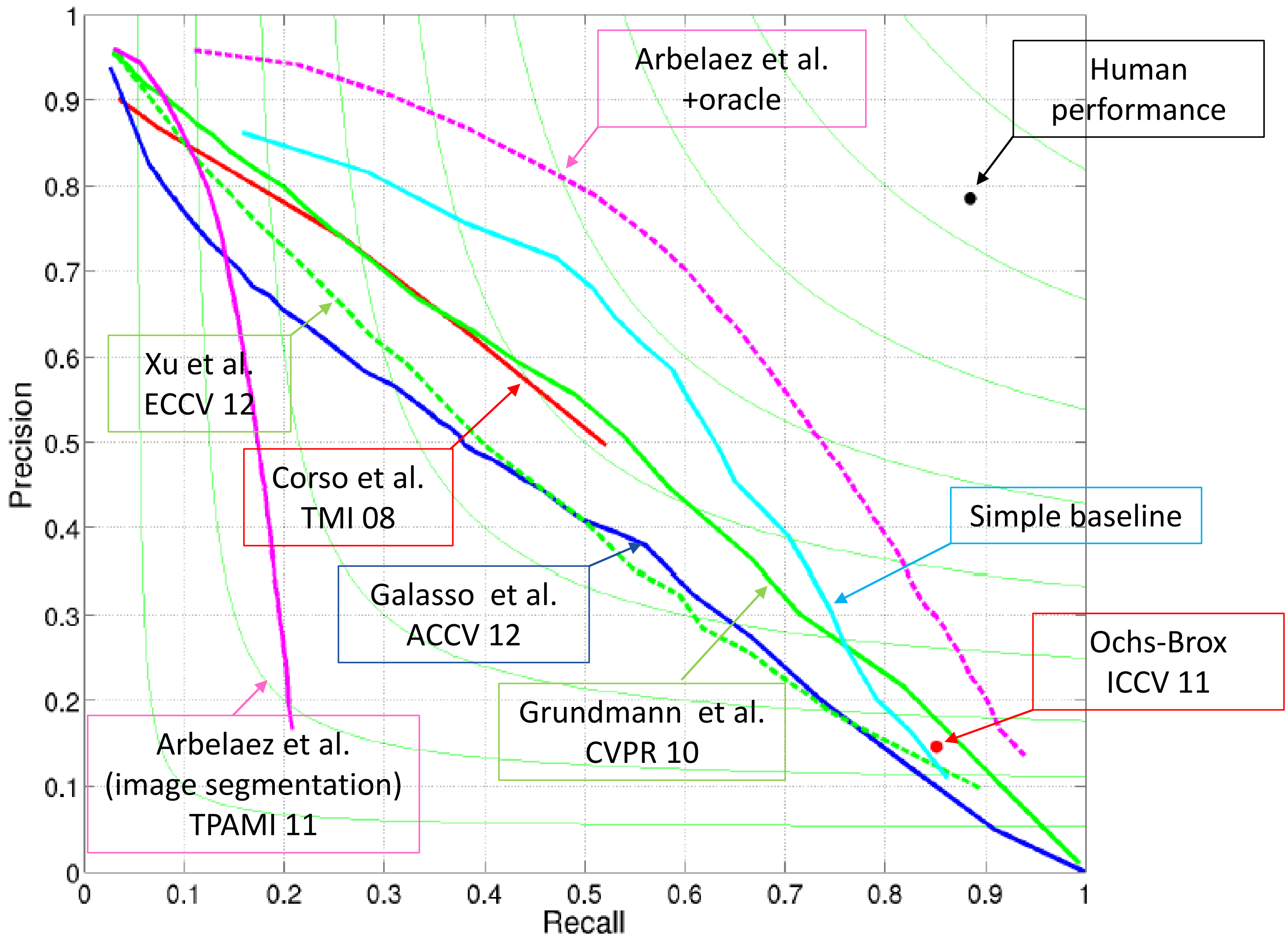
P=0



R=0

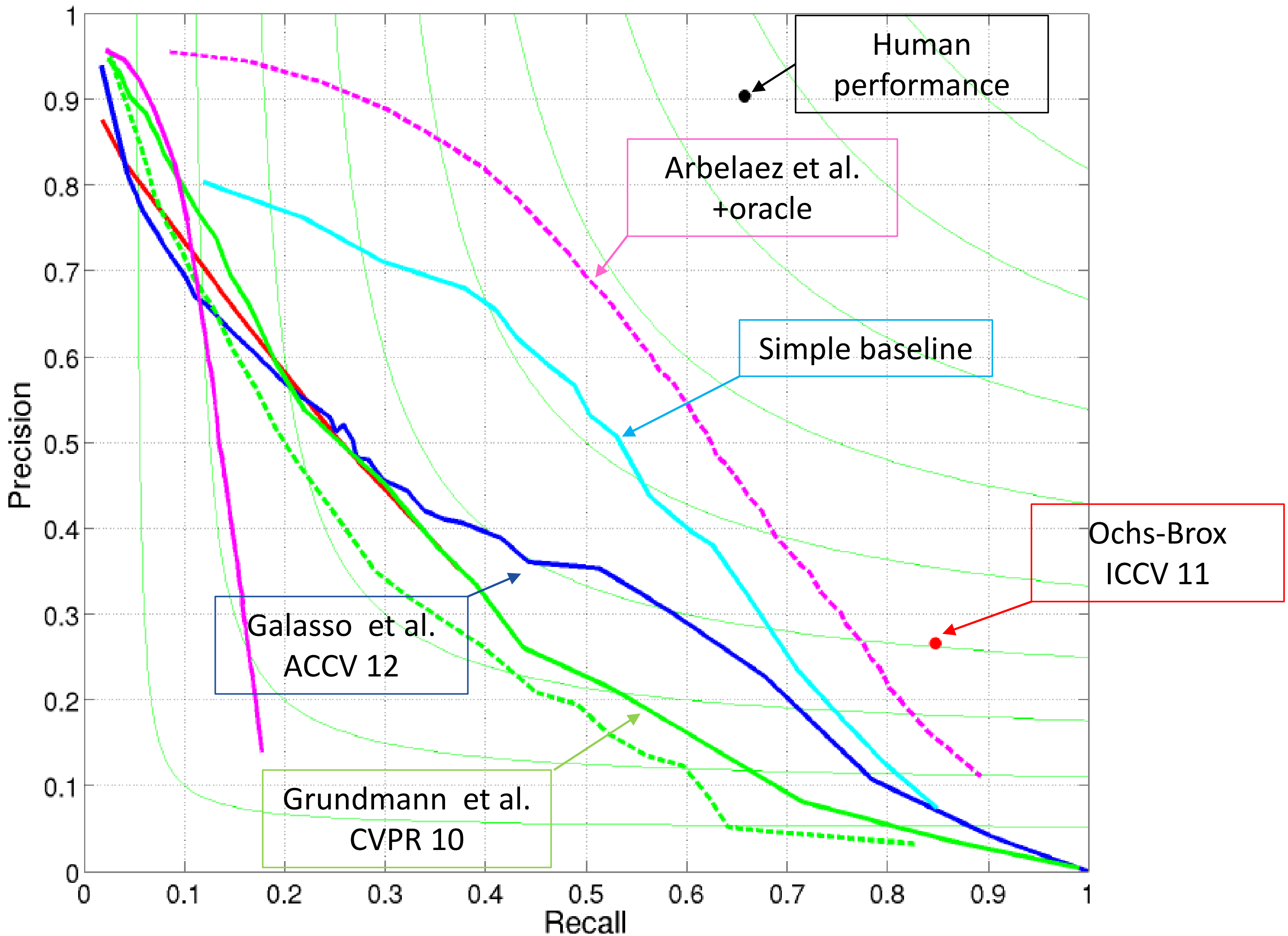
- Many-to-one matching (important for supervoxels)
- Normalization penalizes extreme segmentations

Results


Arbelaez et al.
+oracle
Human
performance
Xu et al.
ECCV 12
Corso et al.
TMI 08
Simple baseline
Galasso et al.
ACCV 12
Ochs-Brox
ICCV 11
Grundmann et al.
CVPR 10
Arbelaez et al.
(image segmentation)
TPAMI 11
Precision
Recall
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Motion segmentation subtask

VSB100



About the “simple baseline”

1. Take superpixel hierarchy from Arbelaez et al.
2. Propagate labels to next frame using optical flow
3. Next frame:
   label determined by voting

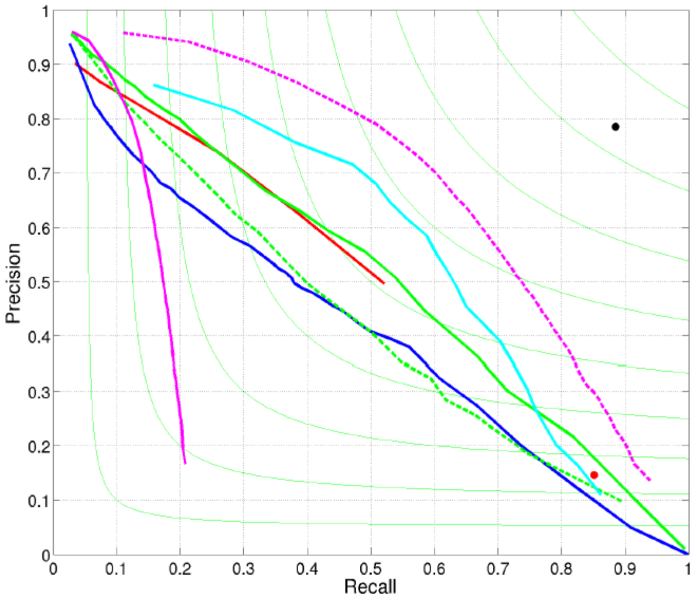
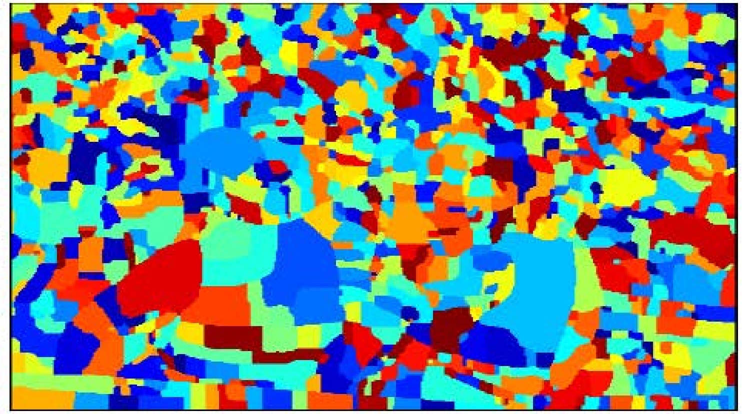


Image segmentation + optical flow < video segmentation

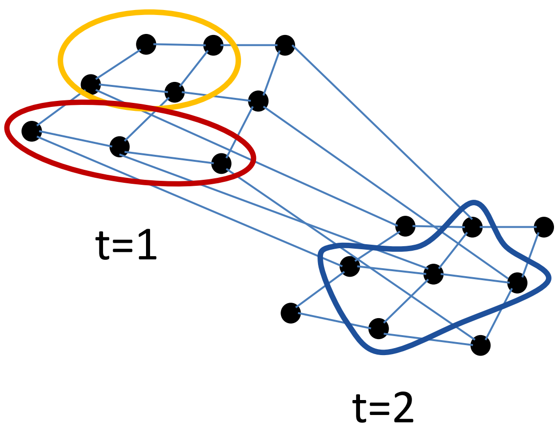
There is work to do

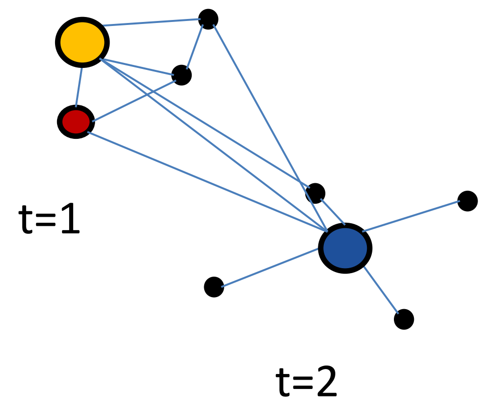
Balanced graph reduction

Original pixels



Superpixels





Fabio Galasso

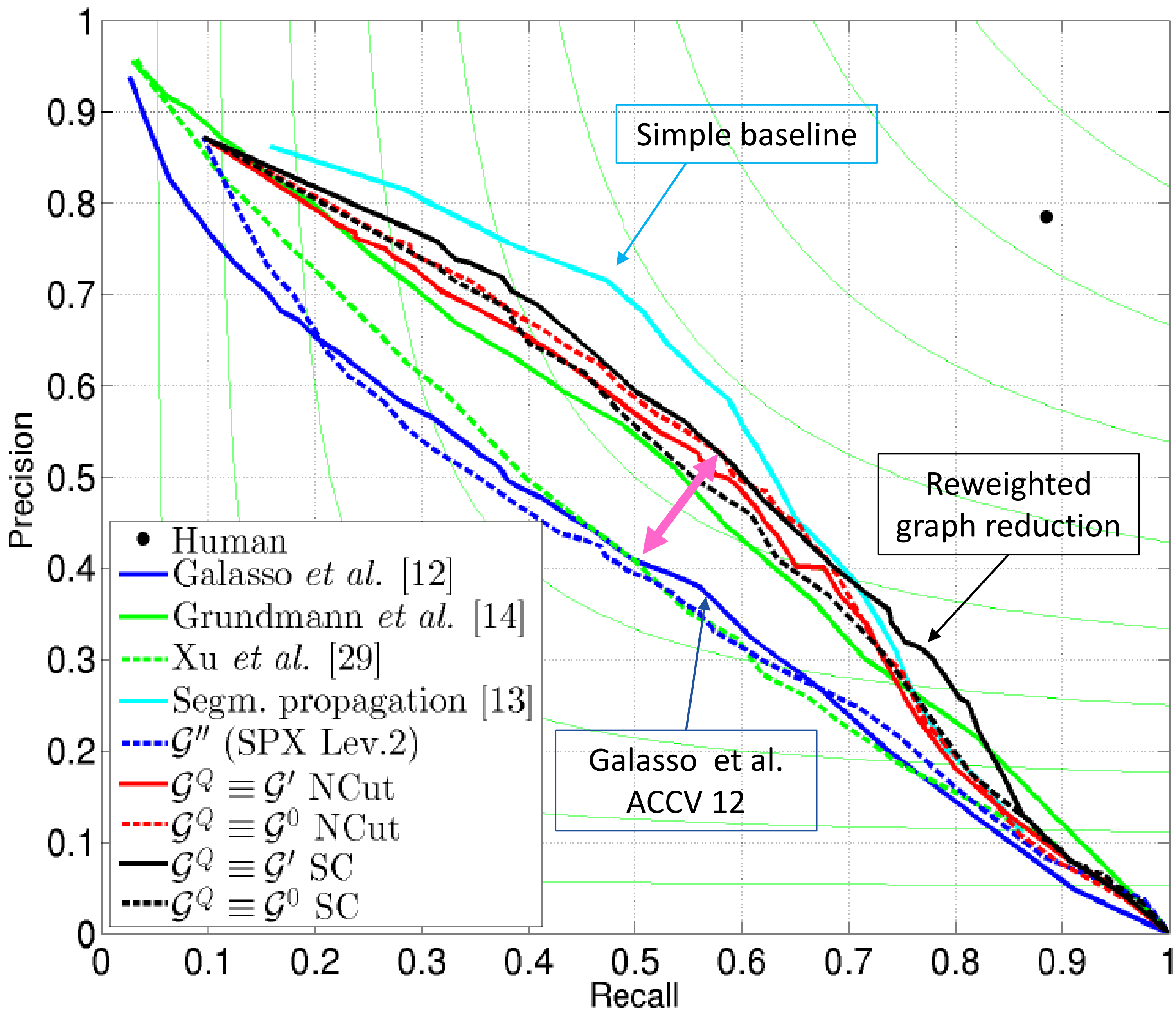
Margret Keuper

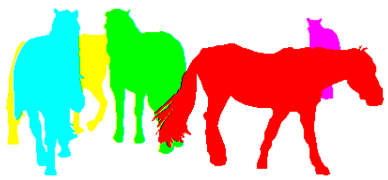
Bernt Schiele

Galasso et al. CVPR 14

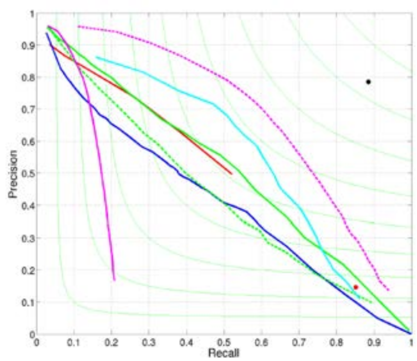
Edge reweighting necessary for weight balancing in spectral clustering

Balancing clearly improves results

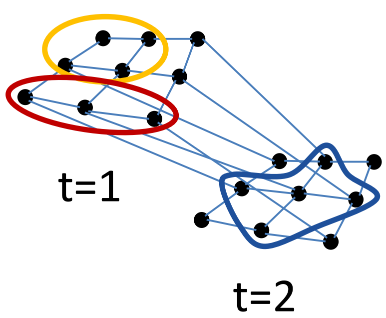




FBMS-59:
Motion segmentation benchmark



VSB-100:
General video segmentation benchmark



Spectral clustering with superpixels:
Don’t forget to rebalance