THE UNIVERSITY OF TEXAS AT AUSTIN

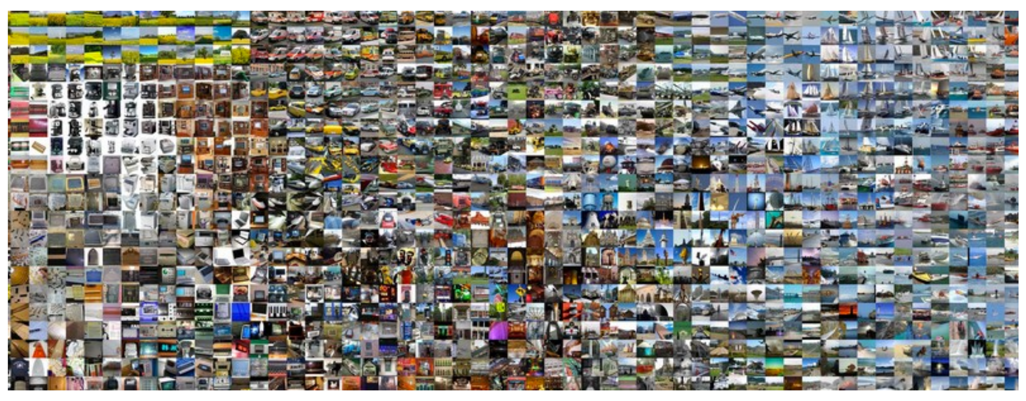
Object-Centric Representation Learning from Unlabeled Videos

Ruohan Gao Dinesh Jayaraman Kristen Grauman
University of Texas at Austin

Problem

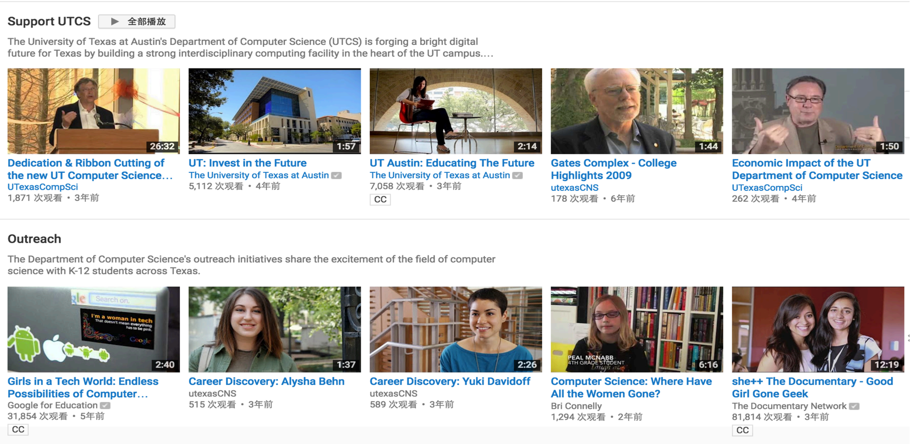
Status quo: Learning from "bags of labeled images"

- ❖ Expensive
- ❖ Limited data
- ❖ Task-specific
- ❖ Not scalable



Solution: Learning unsupervised generic features from unlabeled videos

- ❖ Free
- ❖ Unlimited
- ❖ Generic



Learning from Temporal Coherence

Slow Feature Analysis (SFA):
video frames change slowly over time

Supervision Signal for Feature Learning: Temporally close frames should be close in the deep feature space

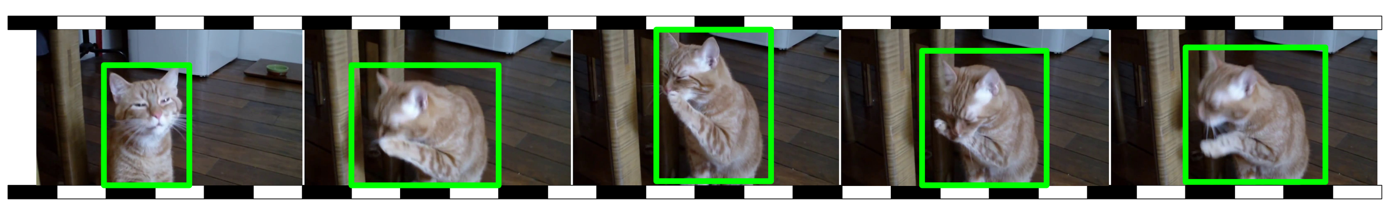
Current Work:

- ❖ **Holistic image embedding:** multiple layers of changes across different regions [Goroshin 2015, Ramanathan 2015, Jayaraman 2016, Mobahi 2009, Bengio 2009, …]
- ❖ **Tracking:** error-prone, biased to moving objects and inefficient [Wang 2015, Zou 2011, Zou 2012]

Temporally Coherent Region Proposals

Our idea: region proposals of temporally close video frames can provide supervision

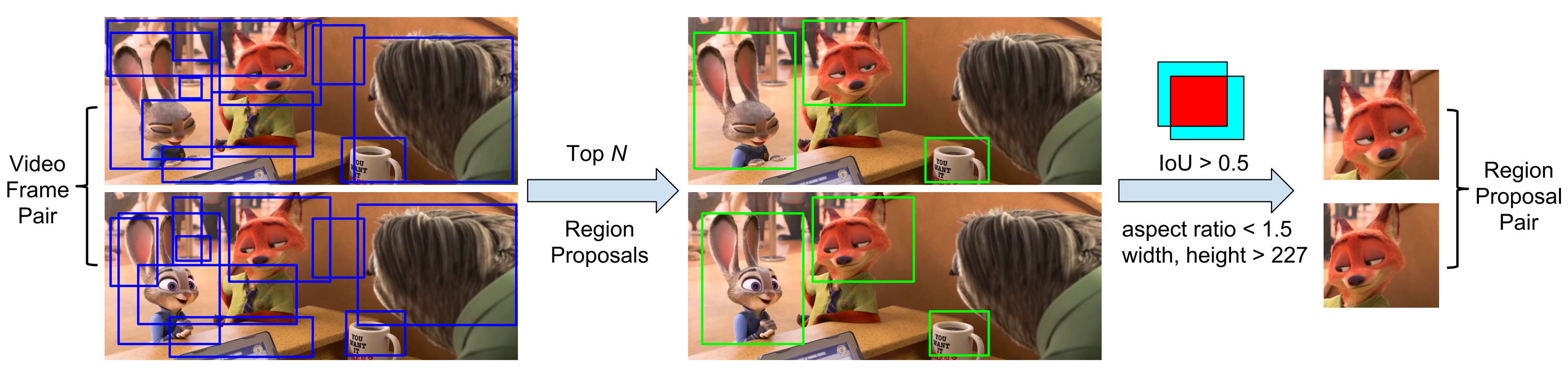
Region Proposals ➡ Selective Search [Uijlings 2013]



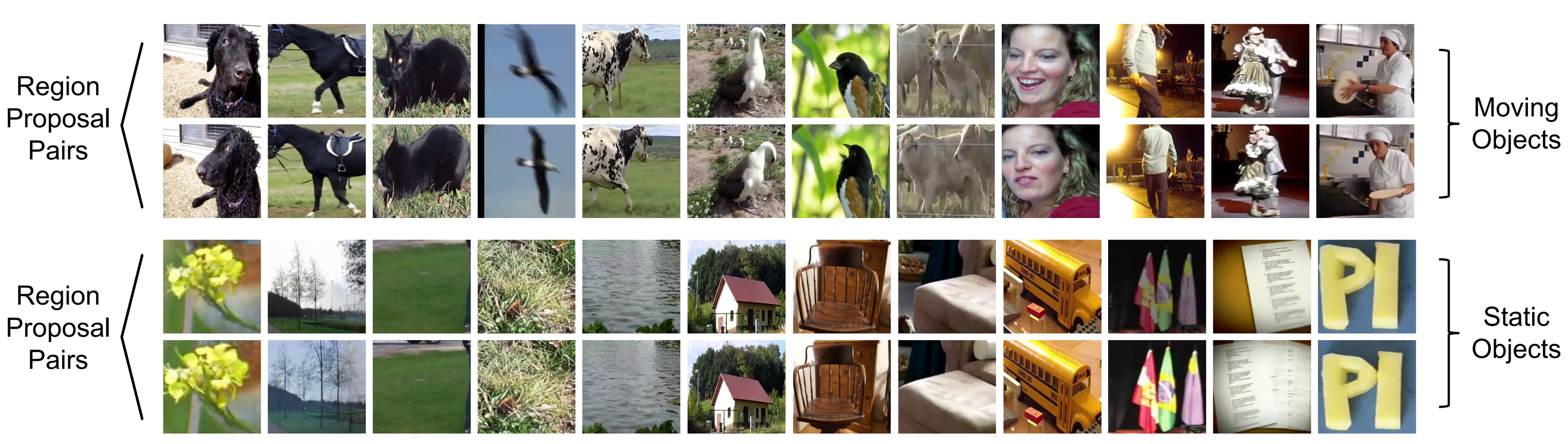
Advantages:

- ❖ capture both static objects and moving objects
- ❖ object-like regions are informative
- ❖ >100 times faster than tracking algorithms

Region Proposal Pair Generation:



Examples of Region Proposal Pairs:

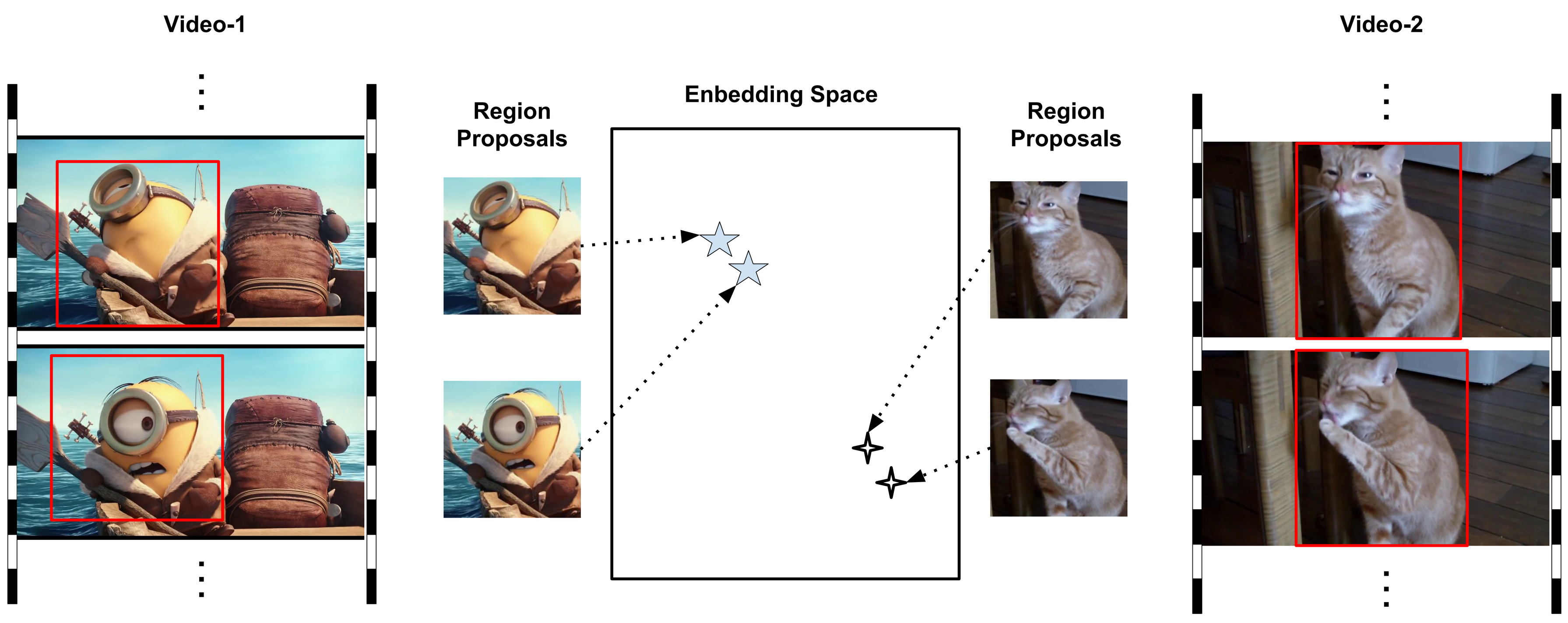


ACCV '16

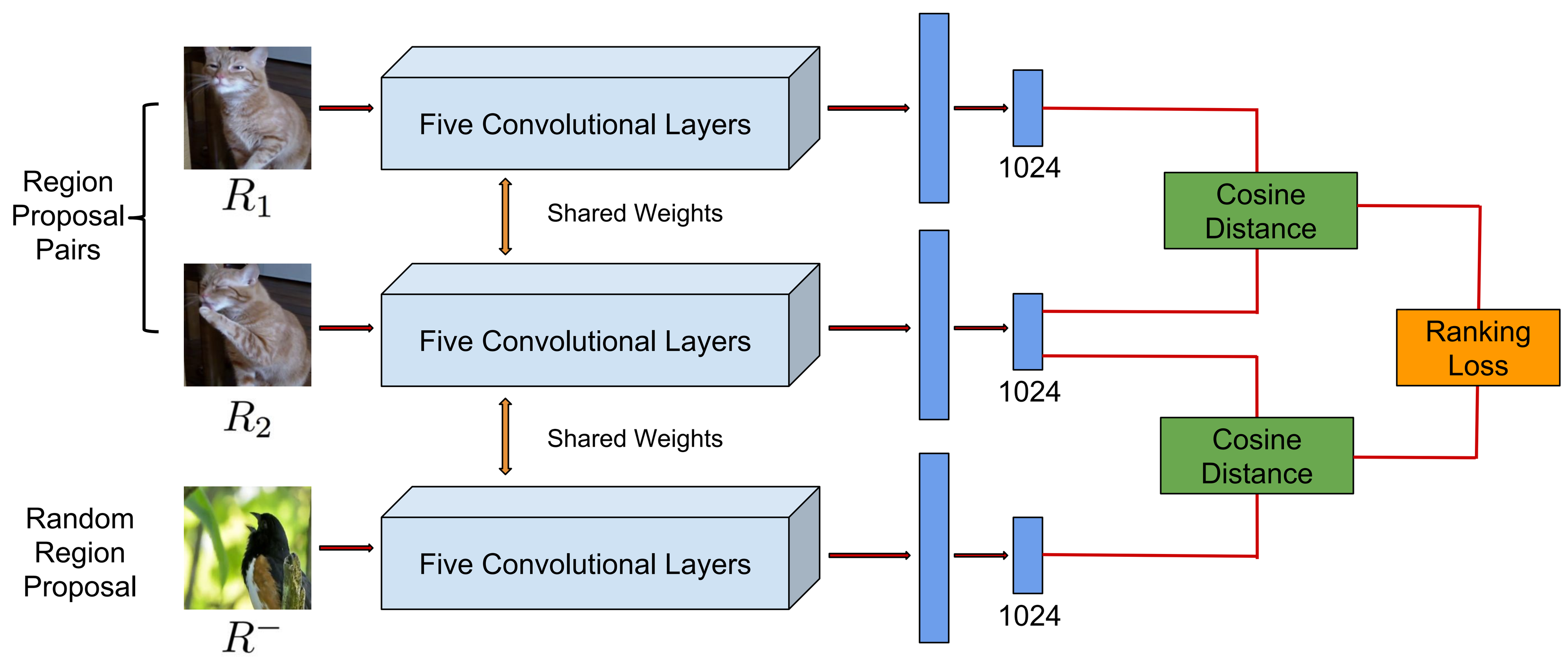
Our pre-trained model is available on our project page: vision.cs.utexas.edu/projects/object_centric_unsup/

Framework

Our Framework: Temporally close region proposals should be close in the deep feature space



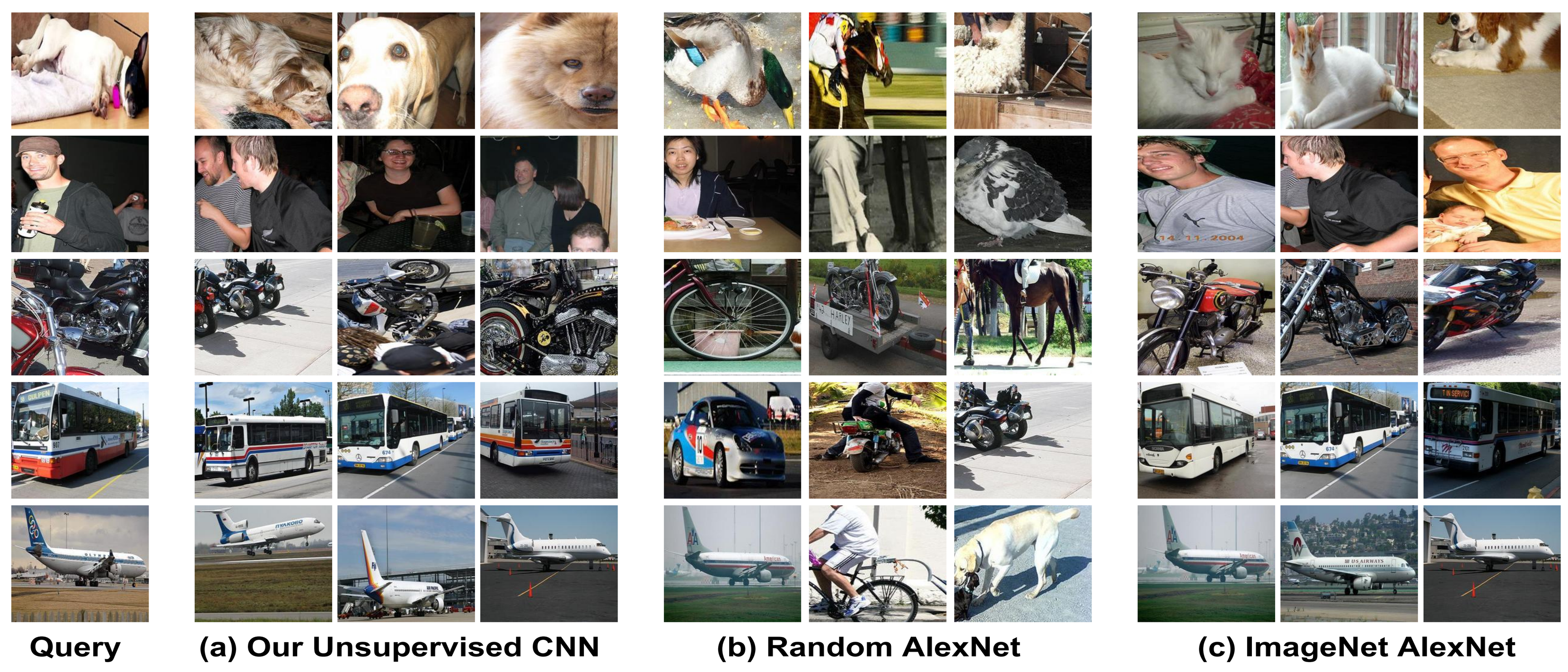
Triplet Embedding: two spatio-temporally close region proposals should be embedded closer than a random region proposal from another different video



Evaluation Results

Data: 25,000 unlabeled videos of various categories from YouTube retrieved based on keywords from VOC

Nearest Neighbor Results: far superior to random AlexNet, and comparable to ImageNet AlexNet



Unsupervised Recognition Results:

Method	Supervision	MIT Indoor 67	VOC 2007	VOC 2012
ImageNet	1.2M labeled images	54%	71%	72%
Wang et al. [7]	4M visual tracking pairs	38%	47%	48%
Jayaraman et al. [14]	egomotion	26%	40%	39%
Agrawal et al. [13]	egomotion	25%	38%	37%
Pathak et al. [12]	spatial context	23%	36%	36%
Full-Frame	1M video frame pairs	27%	40%	40%
Square-Region	1M square region pairs	32%	42%	42%
Visual-Tracking [7]	1M visual tracking pairs	31%	42%	42%
Random Gaussian	-	16%	30%	28%
Ours	1M region proposal pairs	34%	46%	47%

Fine-tuning Recognition Results:

Pretraining Method	Supervision	MIT Indoor 67	VOC 2007	VOC 2012
ImageNet	1.2M labeled images	61.6%	71.1%	70.2%
Wang et al. [7]	4M visual tracking pairs	41.6%	47.8%	47.4%
Jayaraman et al. [14]	egomotion	31.9%	41.7%	40.7%
Agrawal et al. [13]	egomotion	32.7%	42.4%	40.2%
Pathak et al. [12]	spatial context	34.2%	42.7%	41.4%
Full-Frame	1M video frame pairs	33.4%	41.9%	40.3%
Square-Region	1M square region pairs	35.4%	43.2%	42.3%
Visual-Tracking [7]	1M visual tracking pairs	36.6%	43.6%	42.1%
Random Gaussian	-	28.9%	41.3%	39.1%
Ours	1M region proposal pairs	38.1%	45.6%	44.1%