

DEEP LEARNING FOR COMPUTER VISION

Summer Seminar UPC TelecomBCN, 4 - 8 July 2016



Instructors



Xavier
Giró-i-Nieto



Elisa
Sayrol



Amaia
Salvador



Jordi
Torres



Eva
Mohedano



Kevin
McGuinness

Organizers



+ info: TelecomBCN.DeepLearning.Barcelona

Day 4 Lecture 4 Video Analytics



Xavier Giró-i-Nieto



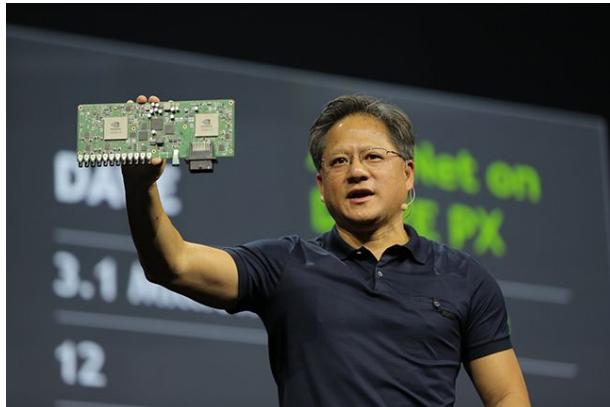
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Department of Signal Theory
and Communications
Image Processing Group

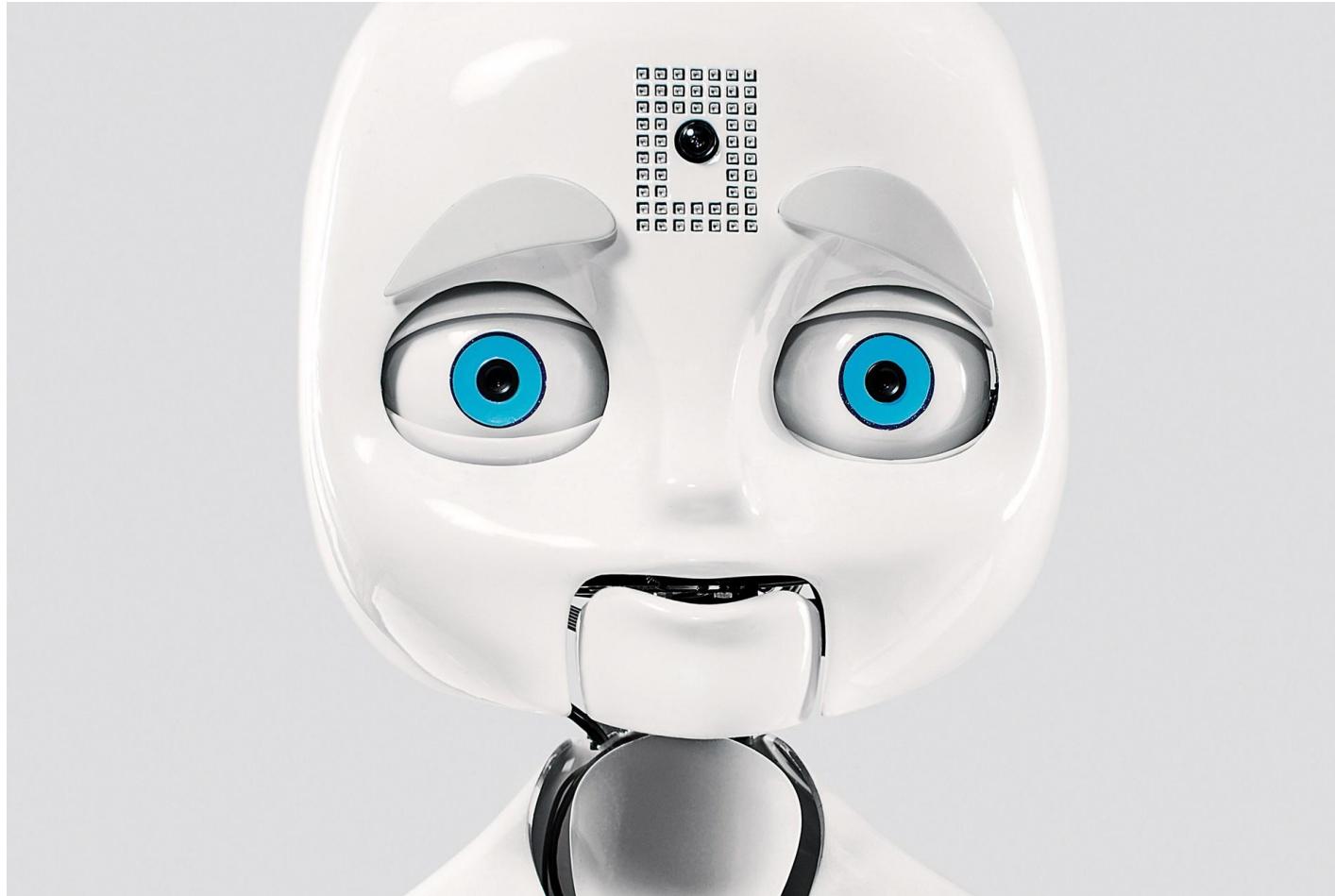
Motivation



Motivation



Motivation



Outline

1. Scene Classification
2. Object Detection & Tracking

Scene Classification



(Slides by Victor Campos) Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014, June). [Large-scale video classification with convolutional neural networks](#). CVPR 2014

Scene Classification

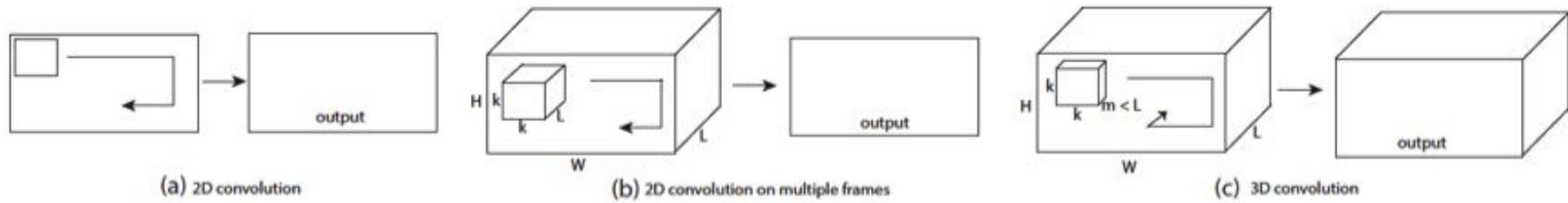
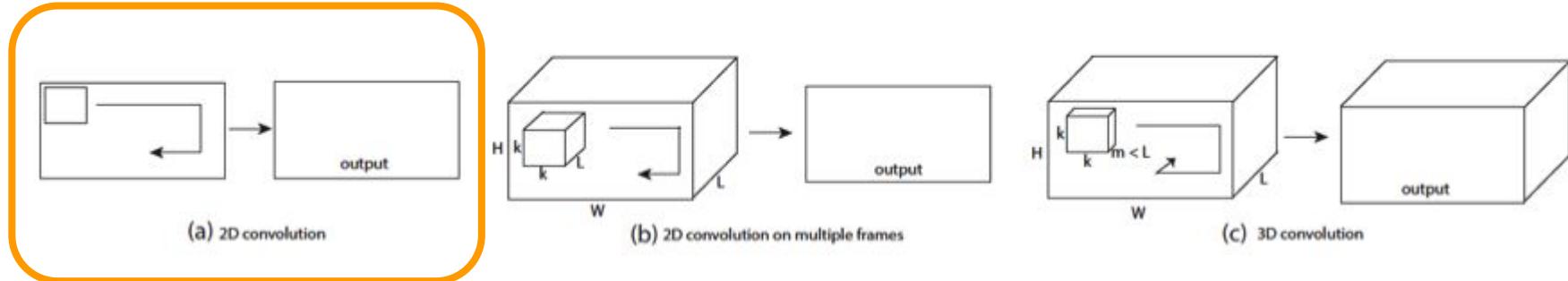


Figure: Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489-4497. 2015

Scene Classification



Previous lectures

Figure: Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489-4497. 2015

Scene Classification

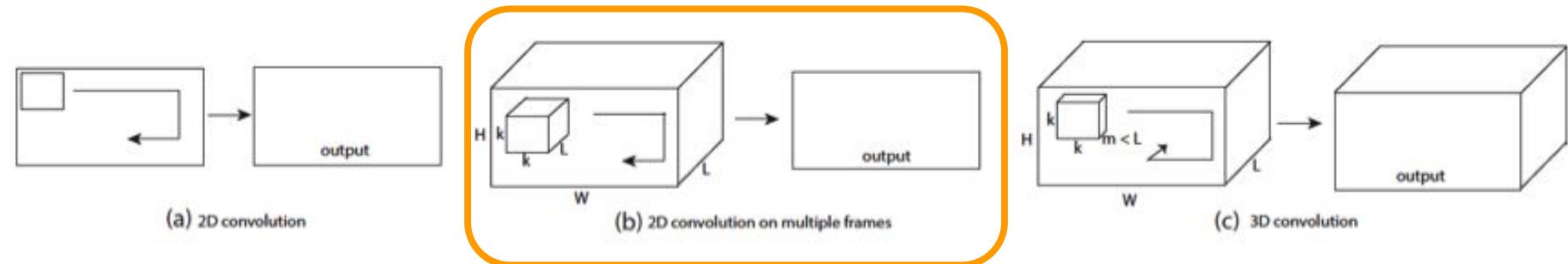
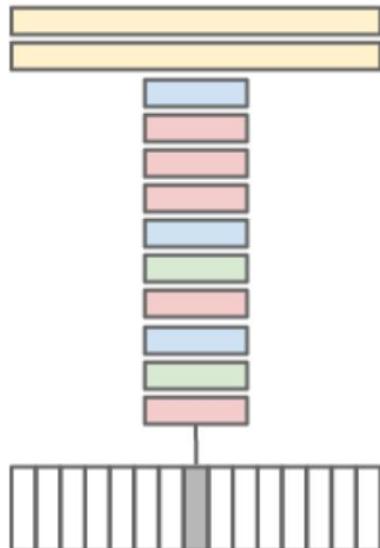


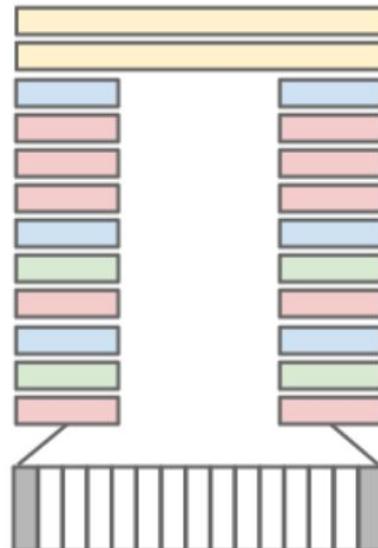
Figure: Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489-4497. 2015

Scene Classification: DeepVideo: Architectures

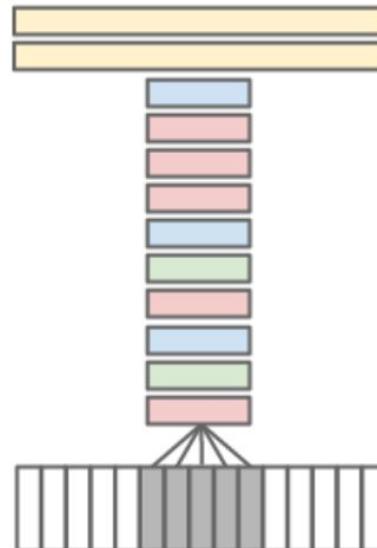
Single Frame



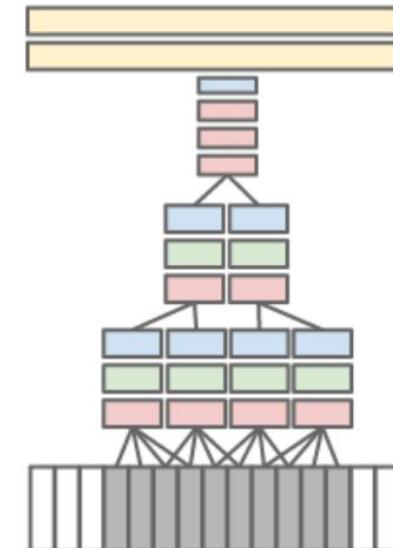
Late Fusion



Early Fusion

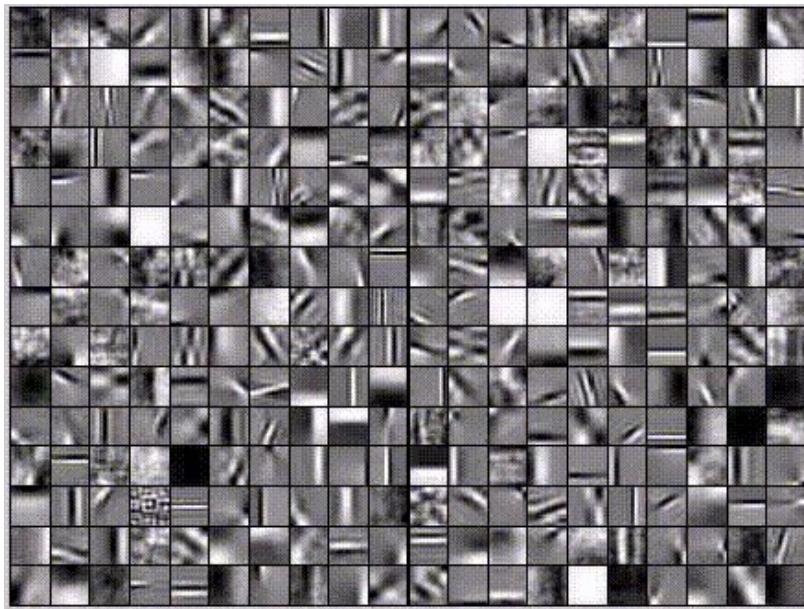


Slow Fusion

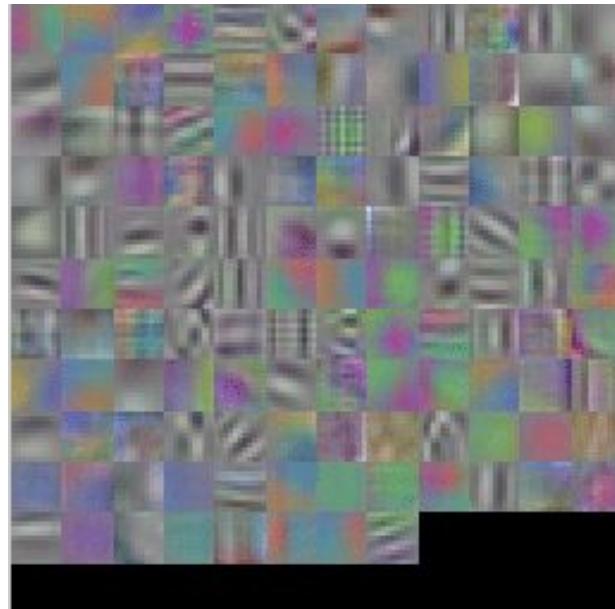


Scene Classification: DeepVideo: Features

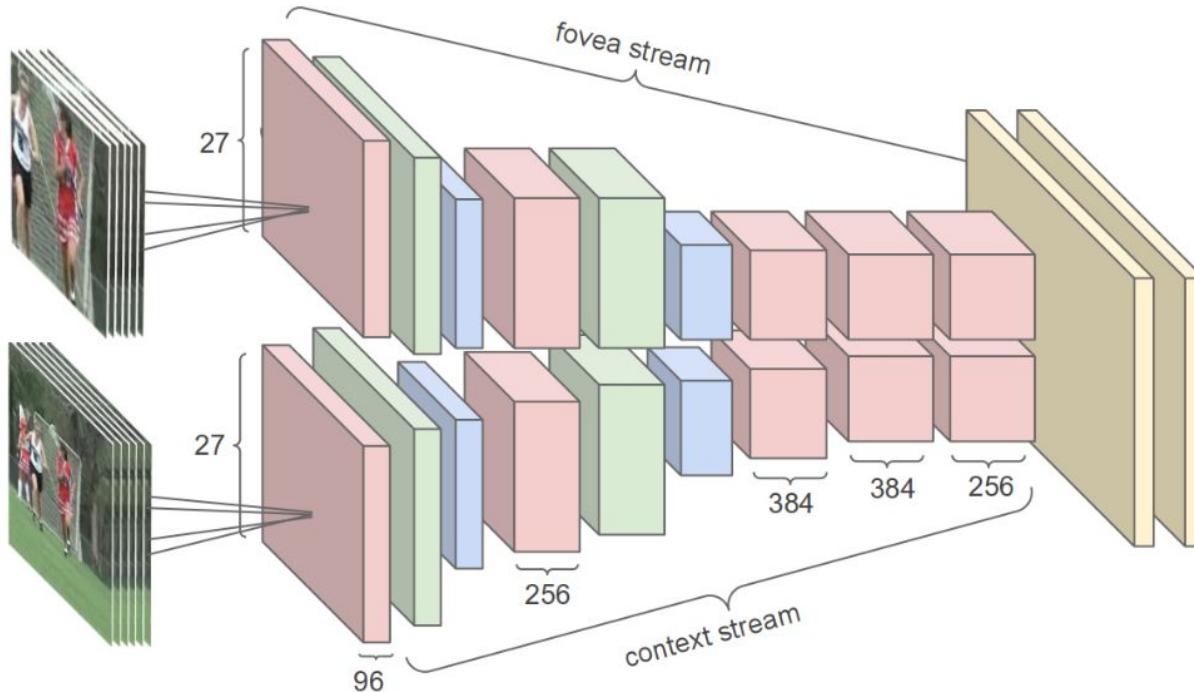
Unsupervised learning [Le et al'11]



Supervised learning [Karpathy et al'14]



Scene Classification: DeepVideo: Multires

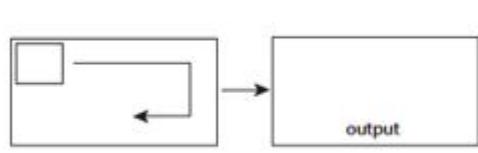


(Slides by Victor Campos) Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014, June). [Large-scale video classification with convolutional neural networks](#). CVPR 2014

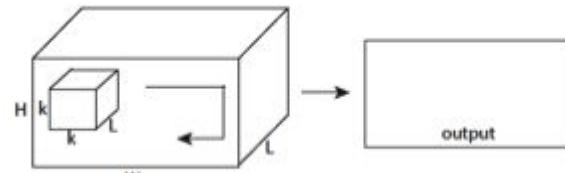
Scene Classification: DeepVideo: Results

Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multires	42.4	60.0	78.5
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	41.9	60.9	80.2
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

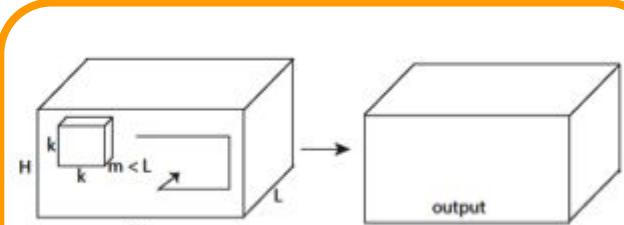
Scene Classification



(a) 2D convolution



(b) 2D convolution on multiple frames



(c) 3D convolution

Figure: Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) CVPR 2015

Scene Classification: C3D

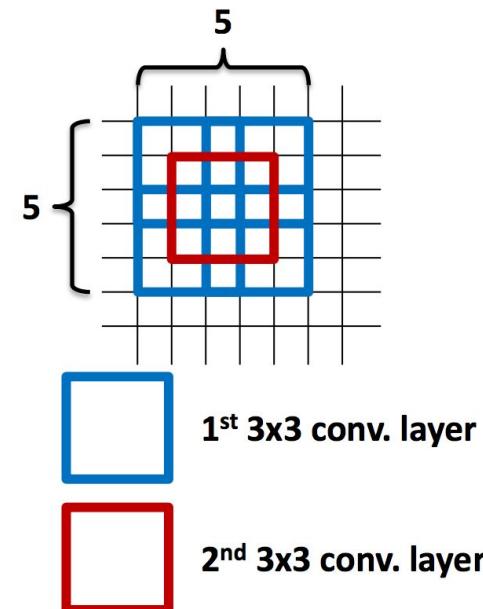


Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#)
CVPR 2015

Scene Classification: C3D: Spatial Dimensions

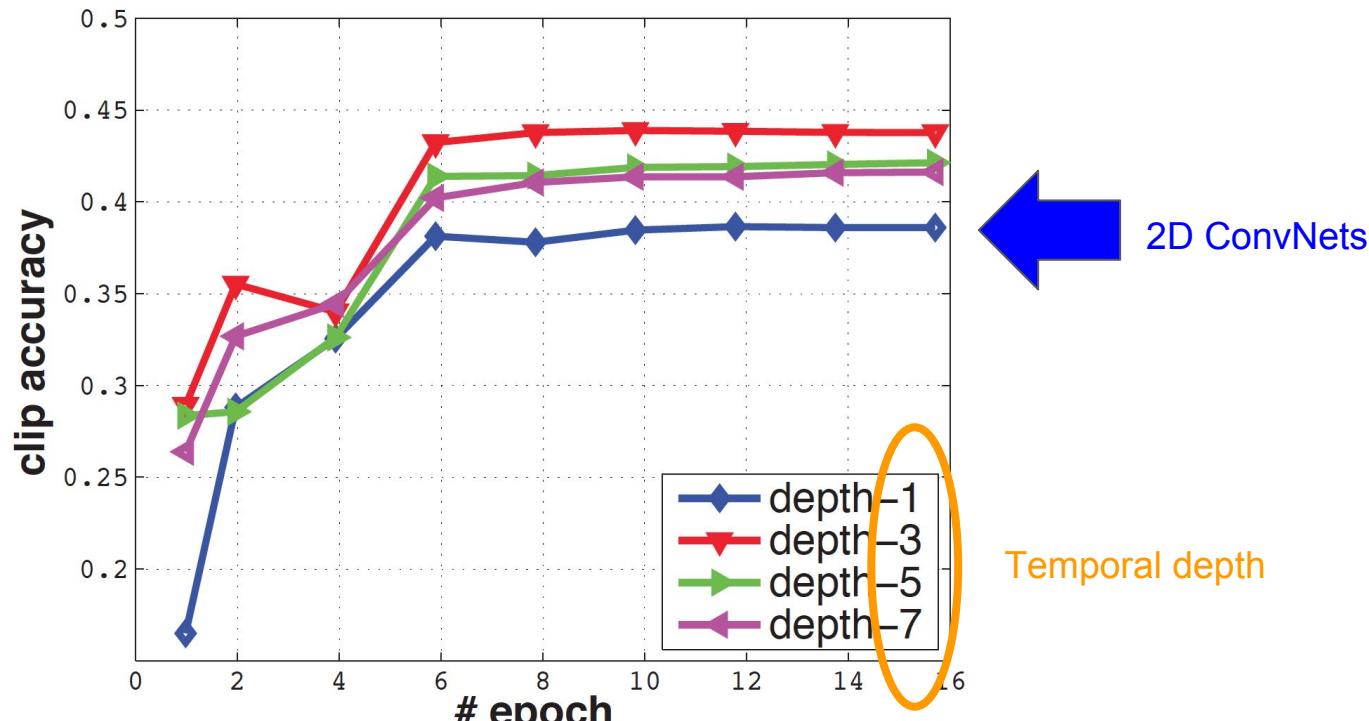
Why 3x3 layers?

- Stacked conv. layers have a large receptive field
 - two 3x3 layers – 5x5 receptive field
 - three 3x3 layers – 7x7 receptive field
- More non-linearity
- Less parameters to learn
 - ~140M per net



Scene Classification: C3D: Temporal dimension

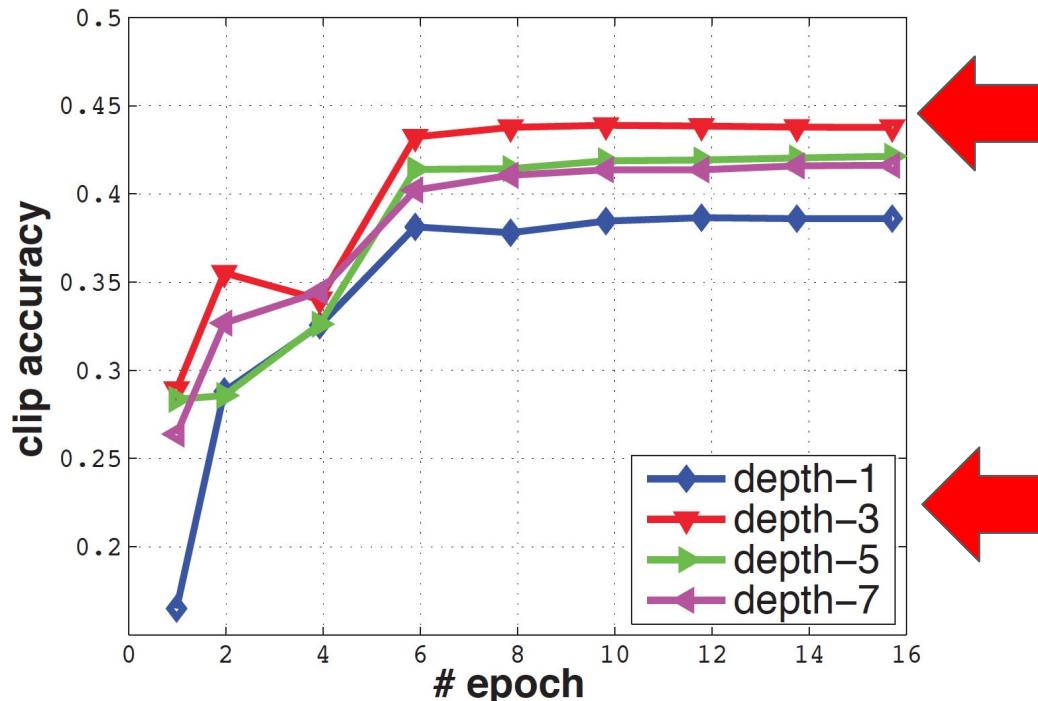
3D ConvNets are more suitable for spatiotemporal feature learning compared to 2D ConvNets



Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) CVPR 2015

Scene Classification: C3D: Temporal dimension

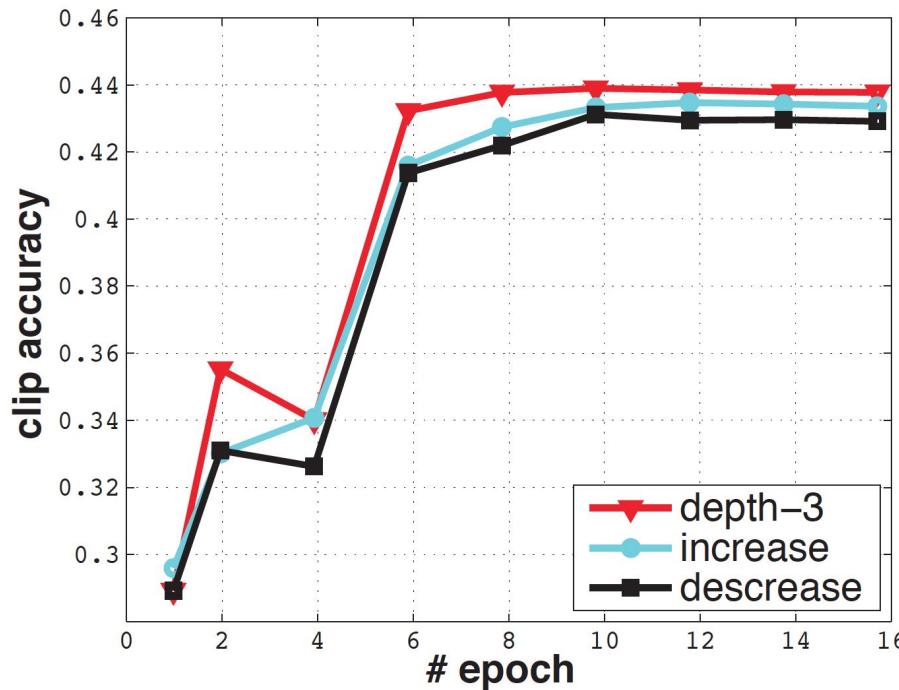
A homogeneous architecture with small $3 \times 3 \times 3$ convolution kernels in all layers is among the best performing architectures for 3D ConvNets



Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) CVPR 2015

Scene Classification: C3D: Temporal dimension

No gain when varying the temporal depth across layers.



Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) CVPR 2015

Scene Classification: C3D: Network Architecture

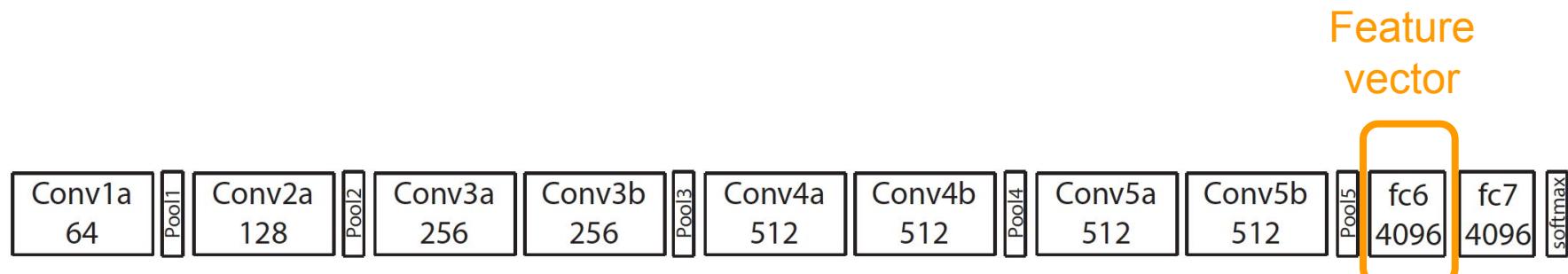
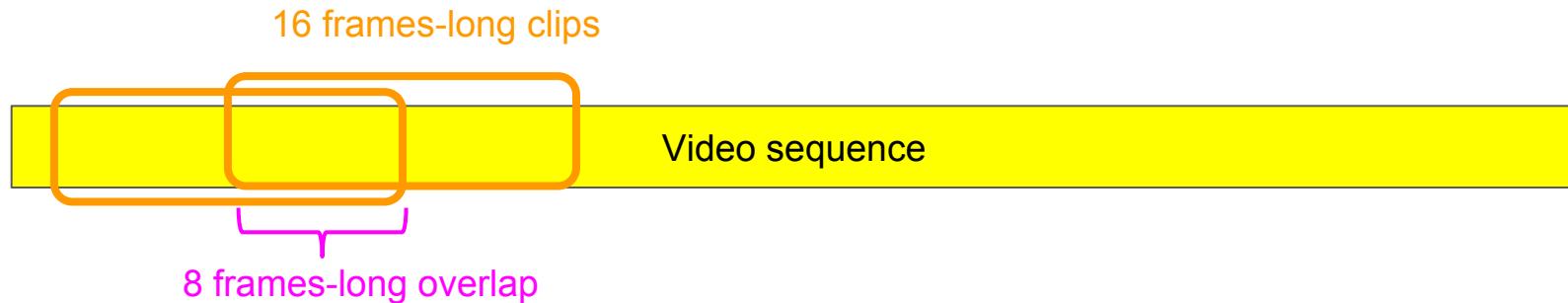


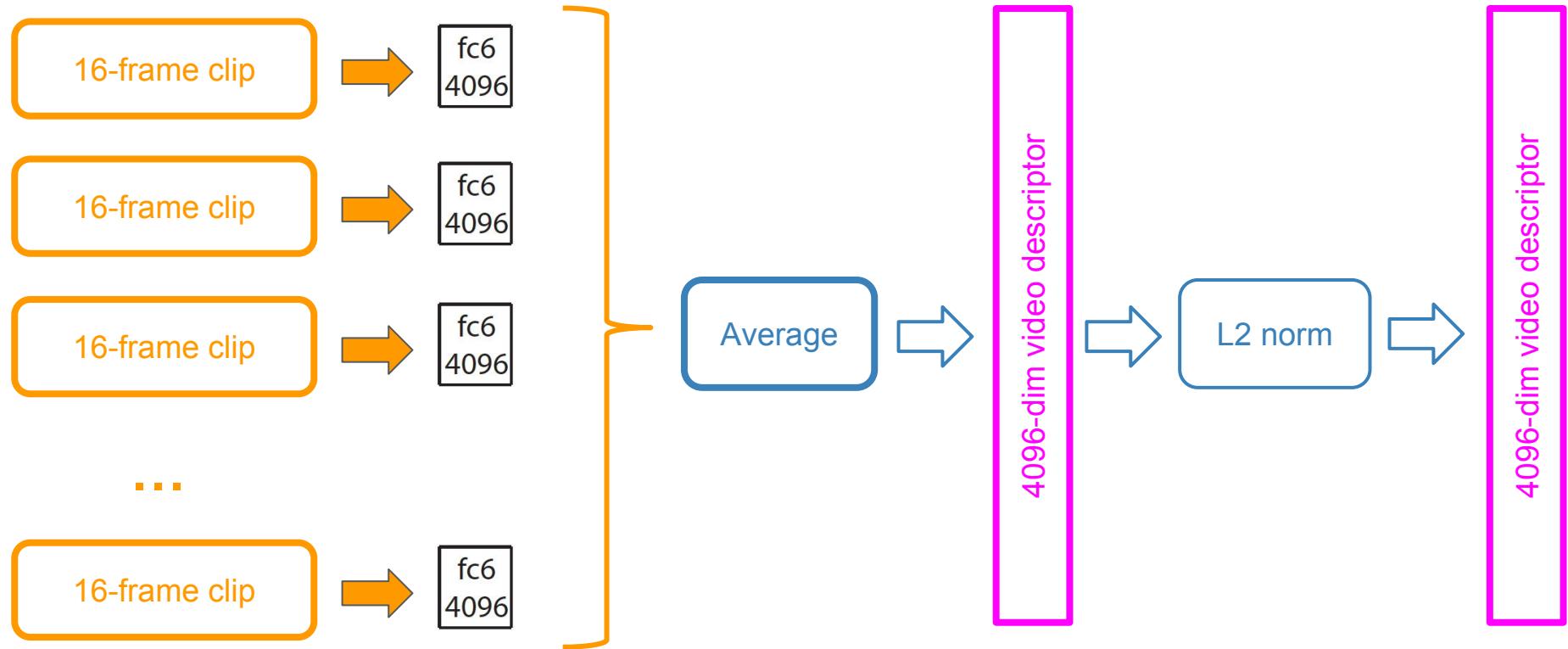
Figure 3. **C3D architecture.** C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from pool1 to pool15. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

Scene Classification: C3D: Feature Vector



Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) CVPR 2015

Scene Classification: C3D: Feature Vector



Tran, Du, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. ["Learning spatiotemporal features with 3D convolutional networks."](#) CVPR 2015

Scene Classification: C3D: Visualization

Based on Deconvnets by [Zeiler and Fergus \[ECCV 2014\]](#) - See [\[ReadCV Slides\]](#) for more details.



Figure 4. **Visualization of C3D model, using the method from [46].** Interestingly, C3D captures appearance for the first few frames but thereafter only attends to salient motion. Best viewed on a color screen.

Scene Classification: C3D: Visualization

C3D + simple linear classifier outperformed state-of-the-art methods on 4 different benchmarks, and were comparable with state of the art methods on other 2 benchmarks

Dataset Task	Sport1M action recognition	UCF101 action recognition	ASLAN action similarity labeling	YUPENN scene classification	UMD scene classification	Object object recognition
Method	[29]	[39]([25])	[31]	[9]	[9]	[32]
Result	90.8	75.8 (89.1)	68.7	96.2	77.7	12.0
C3D	85.2	85.2 (90.4)	78.3	98.1	87.7	22.3

Table 1. **C3D compared to best published results.** C3D outperforms all previous best reported methods on a range of benchmarks except for Sports-1M and UCF101. On UCF101, we report accuracy for two groups of methods. The first set of methods use only RGB frame inputs while the second set of methods (in parentheses) use all possible features (e.g. optical flow, improved Dense Trajectory).

Scene Classification: C3D: Software

Implementation by Michael Gygli (GitHub)

```
In [5]: # Convert the video snippet to the right format
# i.e. (nr in batch, channel, frameNr, y, x) and subtract mean
caffe_snip=c3d.get_snips(snip,image_mean=np.load('snippet_mean.npy'),start=0, with_mirrored=False)

In [6]: # Compile prediction function
prediction = lasagne.layers.get_output(net['prob'], deterministic=True)
pred_fn = theano.function([net['input'].input_var], prediction, allow_input_downcast = True);

In [7]: # Now we can get a prediction
probabilities=pred_fn(caffe_snip).mean(axis=0) # As we average over flipped and non-flipped

In [8]: # Load labels
with open('labels.txt','r') as f:
    class2label=dict(enumerate([name.rstrip('\n') for name in f]))

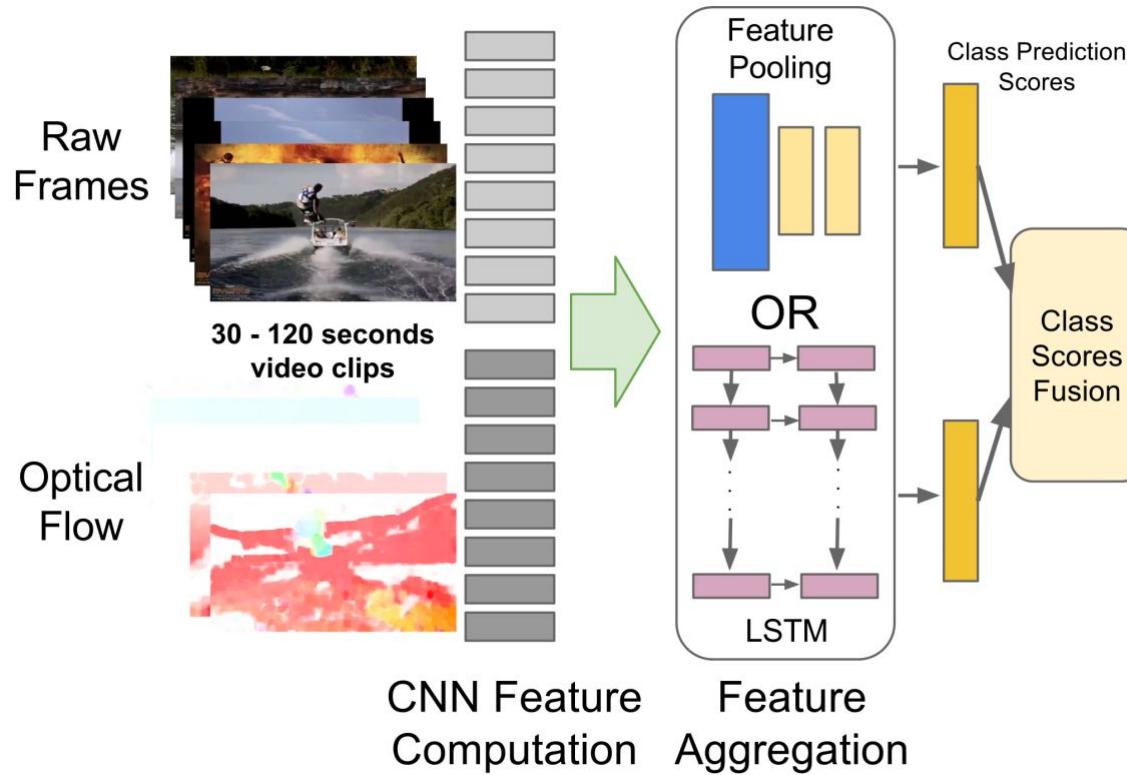
# Show the post probable ones
print('Top 10 class probabilities:')
for class_id in (-probabilities).argsort()[0:10]:
    print('%20s: %.2f%%' % (class2label[class_id],100*probabilities[class_id]))
```

Top 10 class probabilities:

wiffle ball:	29.87%
knife throwing:	13.12%
croquet:	11.36%
disc golf:	5.30%
kickball:	5.15%
rounders:	4.48%
bocce:	3.53%
dodgeball:	2.25%
boomerang:	1.71%
tee ball:	1.39%

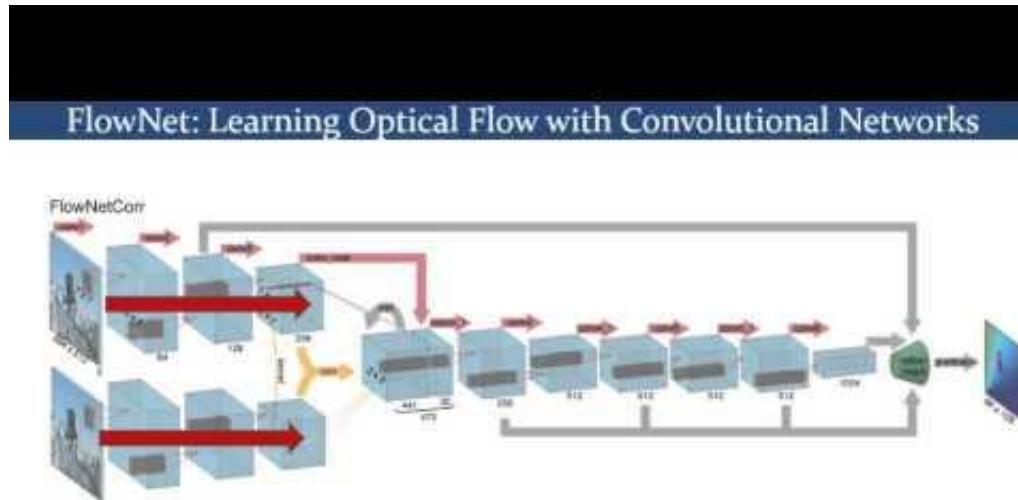


Classification: Image & Optical Flow CNN + LSTM



Yue-Hei Ng, Joe, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. "Beyond short snippets: Deep networks for video classification." CVPR 2015

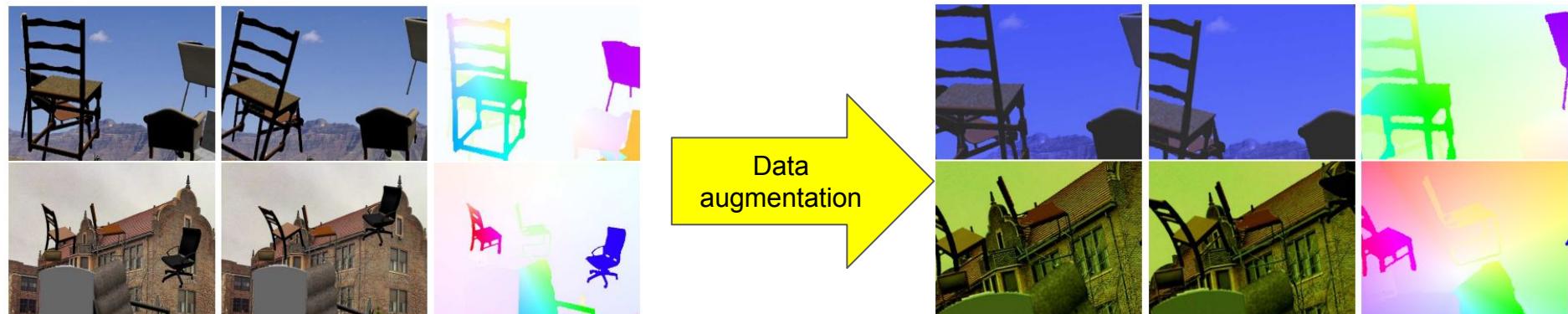
(Scene Classification: Image &) Optical Flow



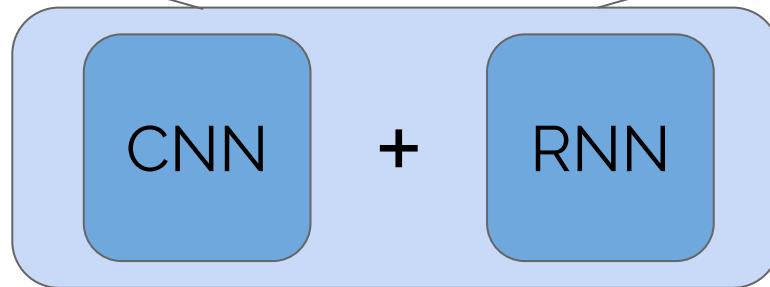
Alternatively, we first process the images separately, then correlate their features at different locations and process further.

(Scene Classification: Image &) Optical Flow

Since existing ground truth datasets are not sufficiently large to train a Convnet, a **synthetic dataset** is generated... and augmented (translation, rotation, scaling transformations; additive Gaussian noise; changes in brightness, contrast, gamma and color).



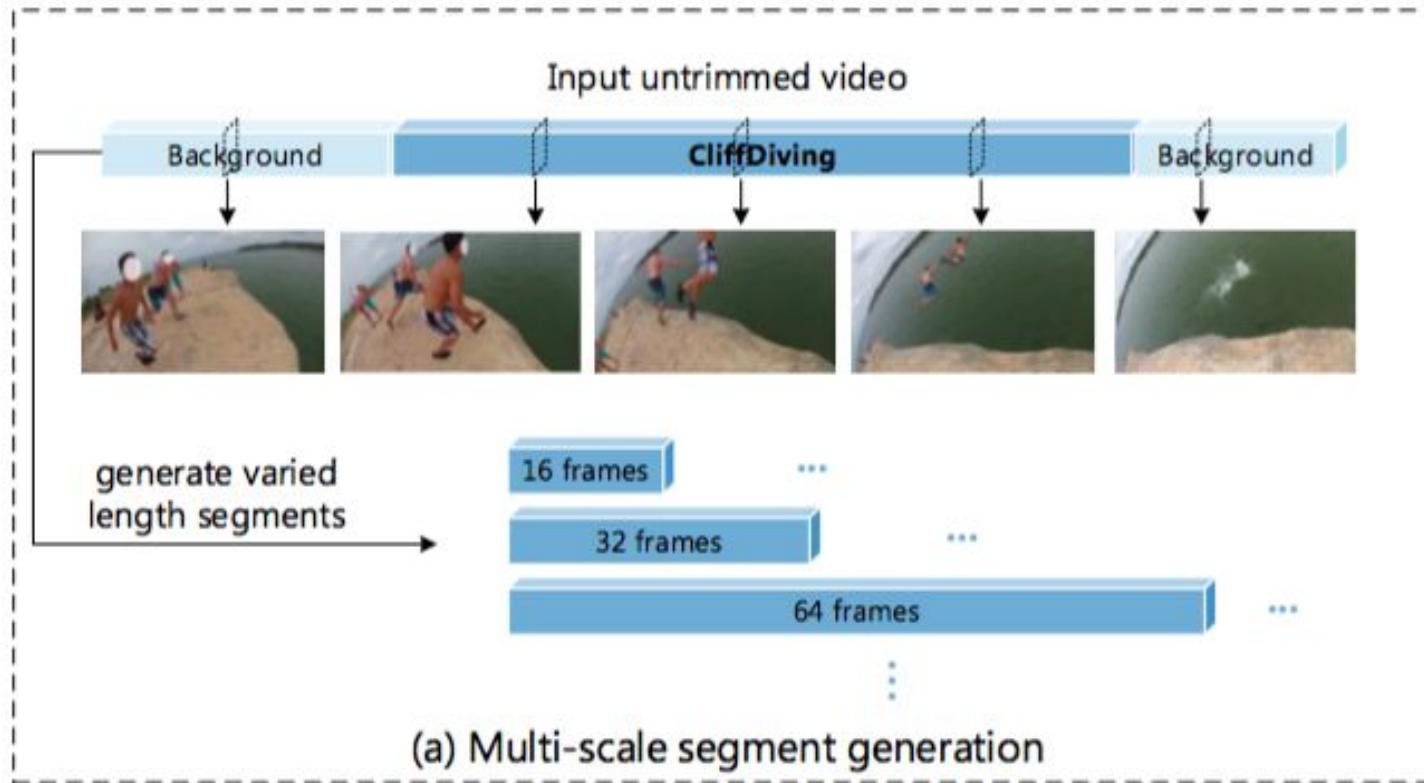
Scene Classification & Detection



“Biking”



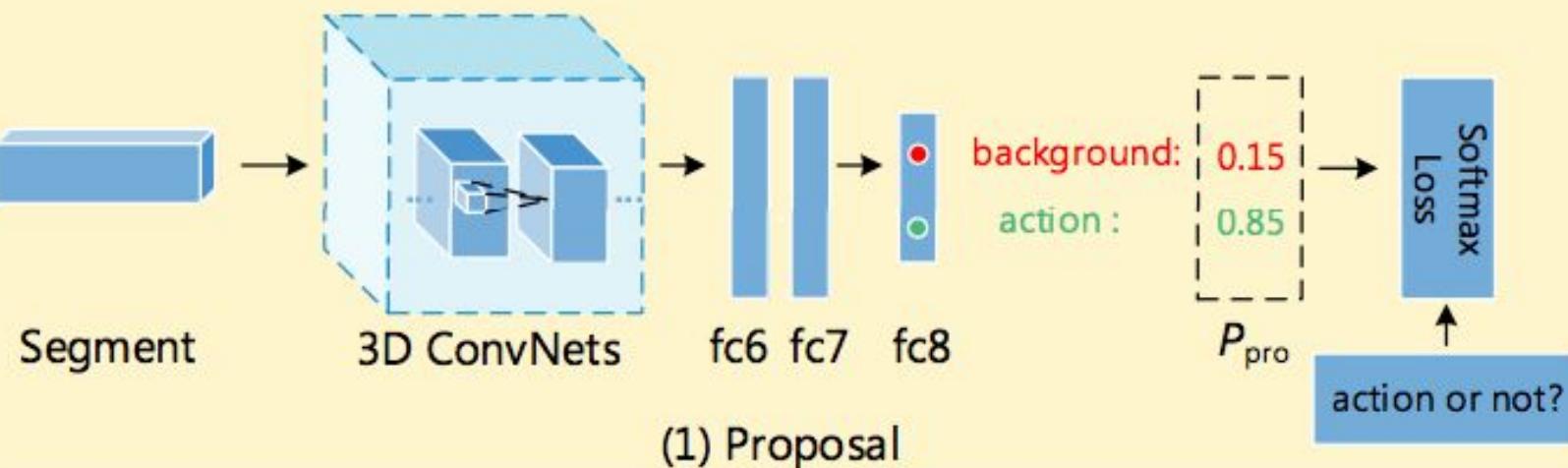
Classification & Detection: Proposals + C3D



(Slidecast and Slides by Alberto Montes) Shou, Zheng, Dongang Wang, and Shih-Fu Chang. "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs." CVPR 2016 [code]

Classification & Detection: Proposals + C3D

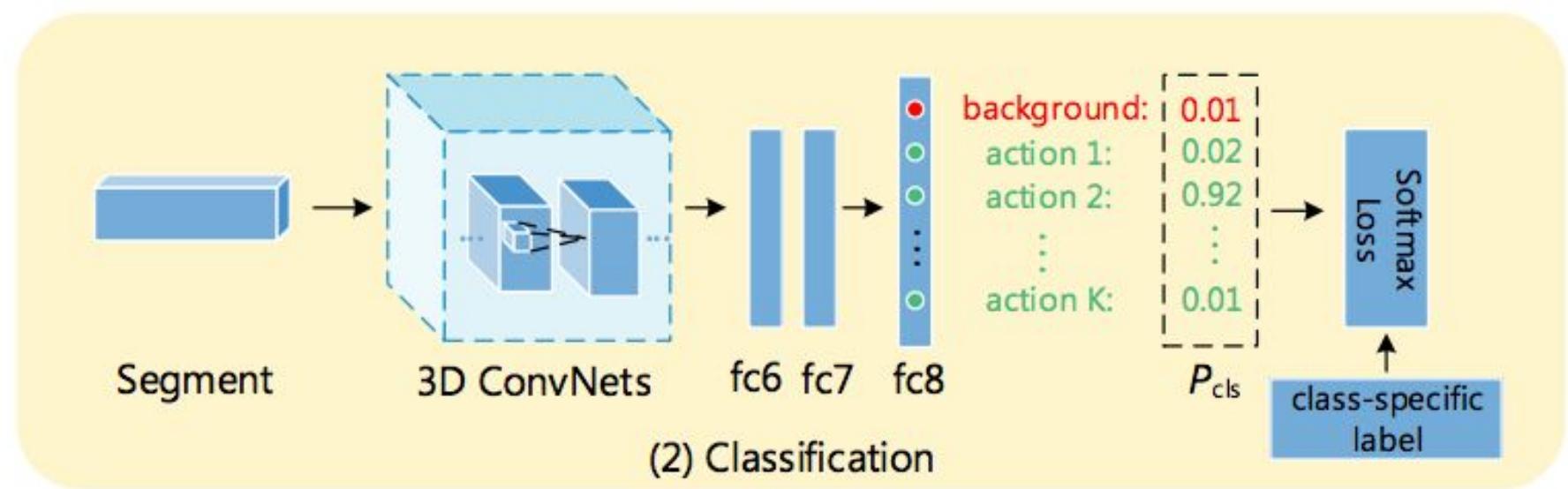
(1) Binary classification: Action or No Action



(Slidecast and Slides by Alberto Montes) Shou, Zheng, Dongang Wang, and Shih-Fu Chang. "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs." CVPR 2016 [code]

Classification & Detection: Proposals + C3D

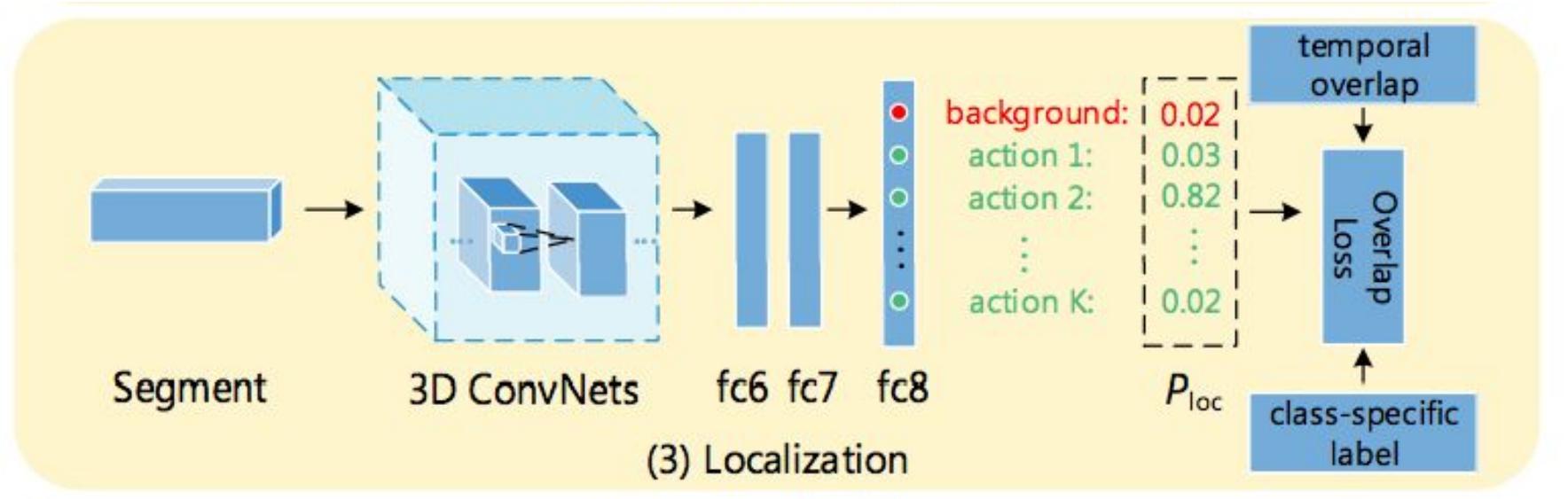
(2) One-vs-all Action classification



(Slidecast and Slides by Alberto Montes) Shou, Zheng, Dongang Wang, and Shih-Fu Chang. "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs." CVPR 2016 [code]

Classification & Detection: Proposals + C3D

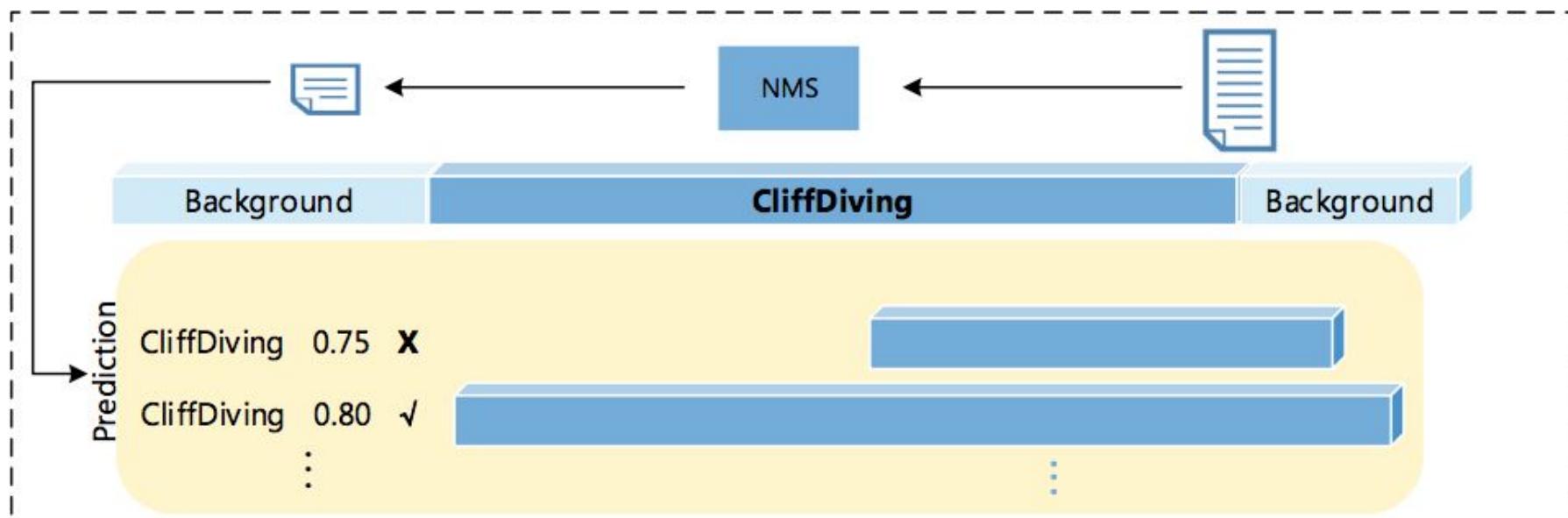
(3) Refinement with temporal-aware loss function



(Slidecast and Slides by Alberto Montes) Shou, Zheng, Dongang Wang, and Shih-Fu Chang. "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs." CVPR 2016 [code]

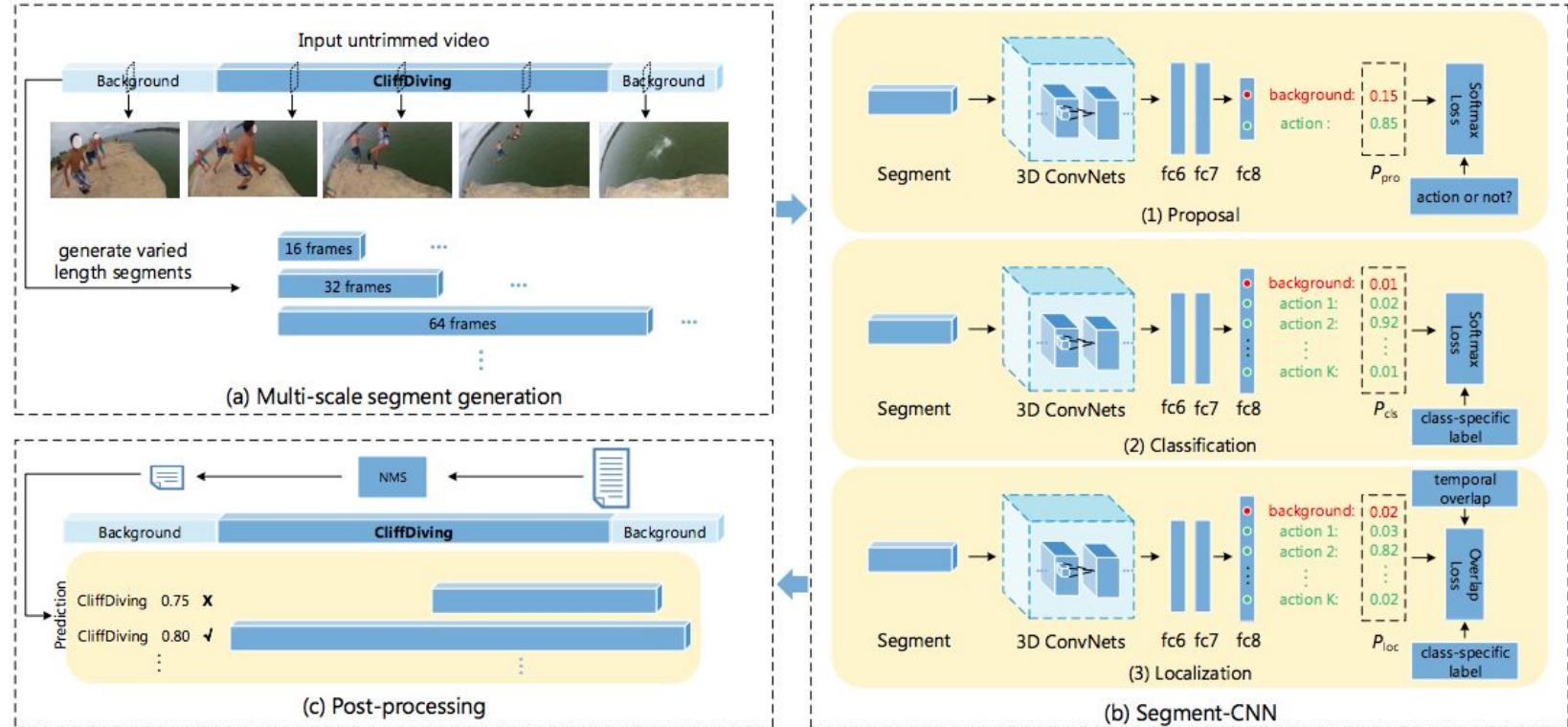
Classification & Detection: Proposals + C3D

Post-processing



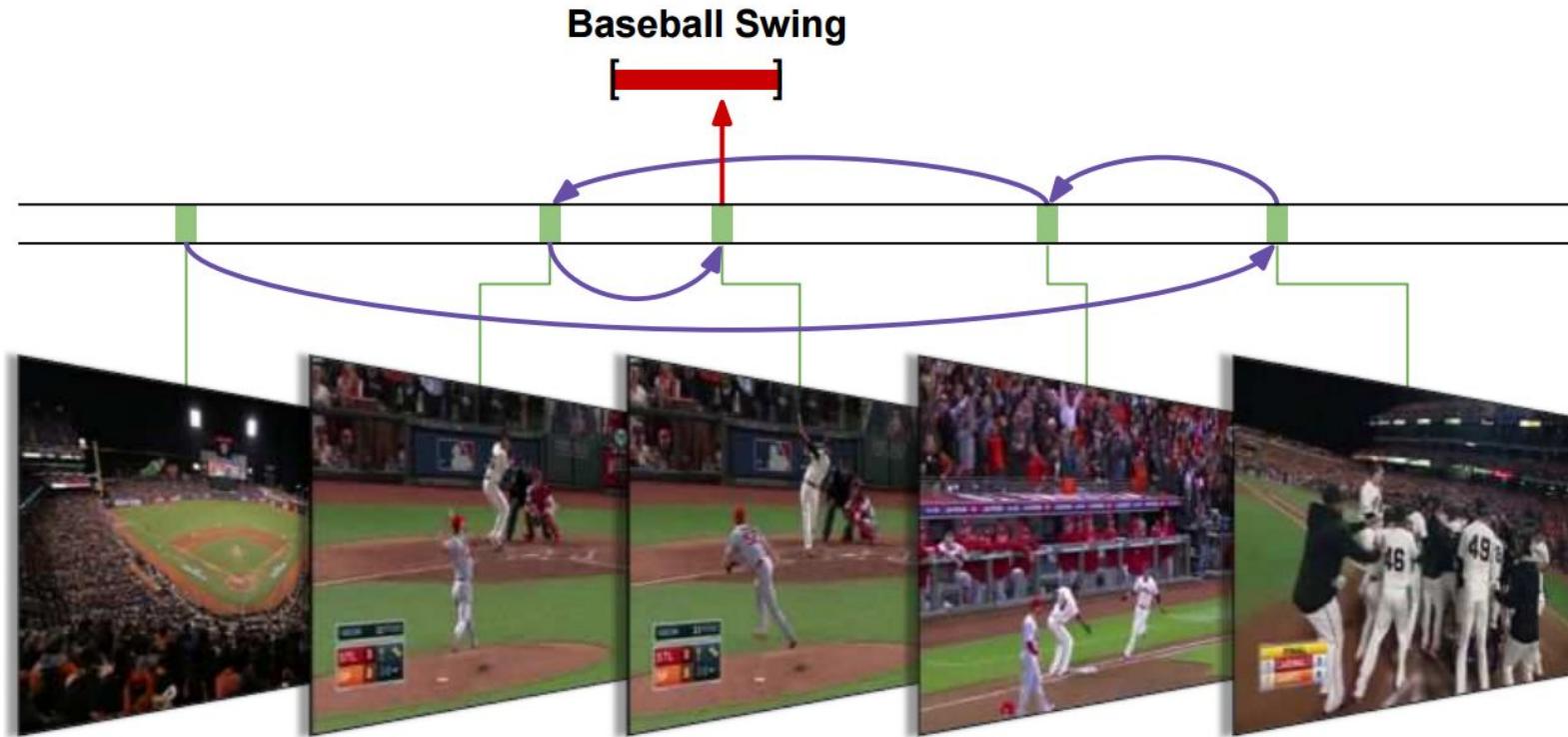
(Slidecast and Slides by Alberto Montes) Shou, Zheng, Dongang Wang, and Shih-Fu Chang. "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs." CVPR 2016 [code]

Classification & Detection: Proposals + C3D



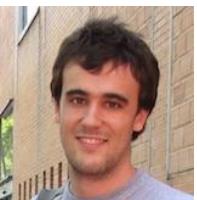
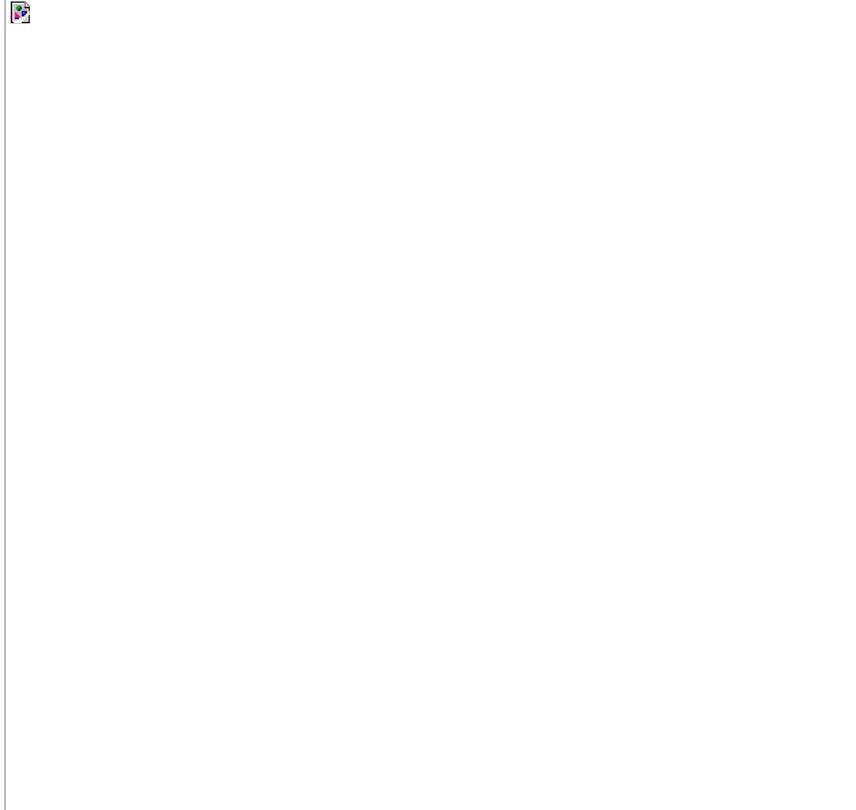
(Slidecast and Slides by Alberto Montes) Shou, Zheng, Dongang Wang, and Shih-Fu Chang. "Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs." CVPR 2016 [code]

Classification & Detection: Image + RNN + Reinforce



Yeung, Serena, Olga Russakovsky, Greg Mori, and Li Fei-Fei. "[End-to-end Learning of Action Detection from Frame Glimpses in Videos.](#)" CVPR 2016

Scene Classification & Detection: C3D + LSTM



Montes A. "[Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks](#)". BSc thesis submitted to ETSETB (2016) [\[code available in Keras\]](#)

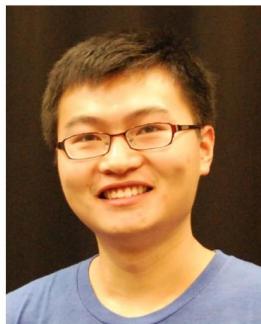
Outline

1. Scene Classification
2. Object Detection & Tracking

Objects: ImageNet Video

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2015

Object Detection from Video (VID)



Wei Liu
UNC Chapel Hill



Olga Russakovsky
CMU



Jia Deng
Univ. of Michigan



Fei-Fei Li
Stanford



Alex Berg
UNC Chapel Hill

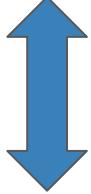
Objects: ImageNet Video

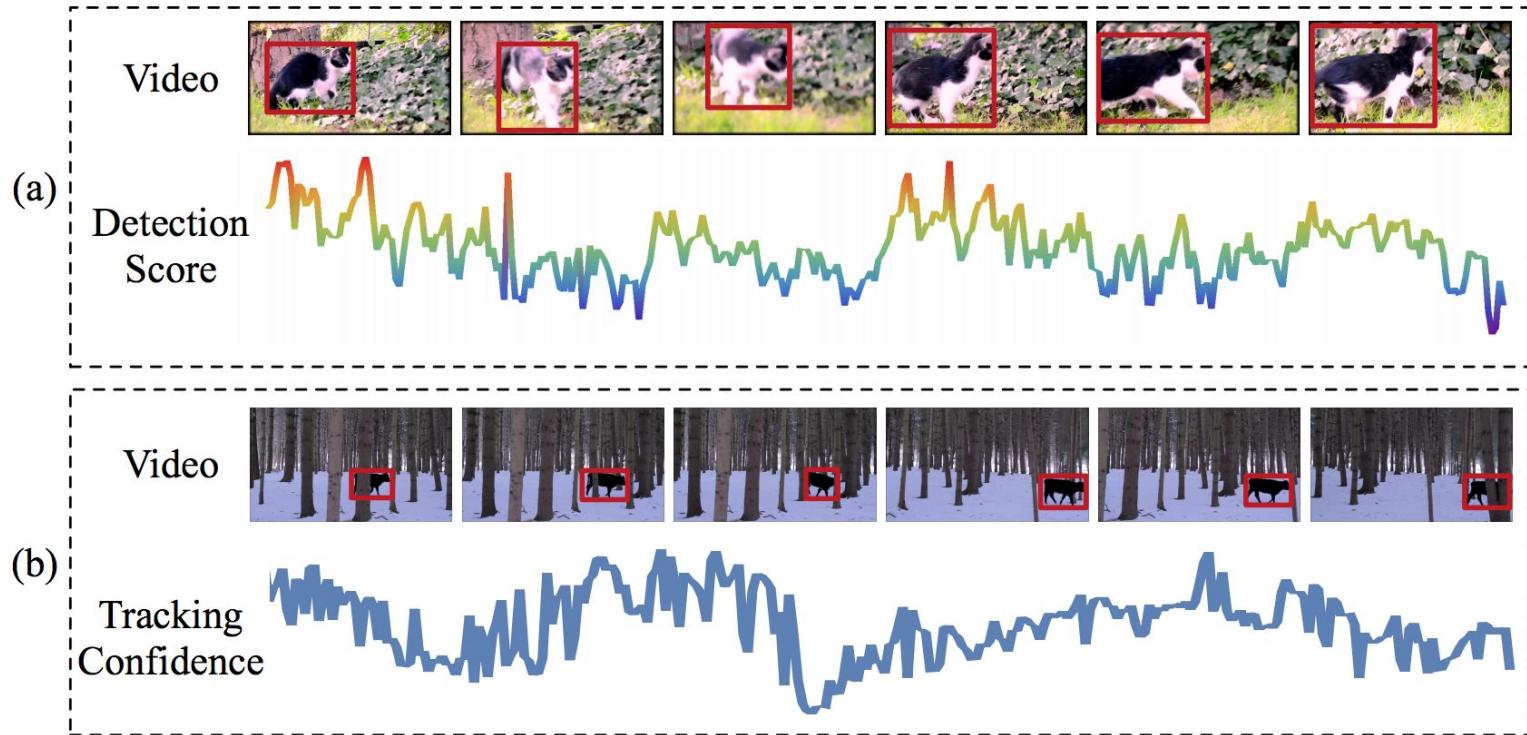
Fully annotated 30 object classes across 5,354 snippets



Allows evaluation of generic object detection
in cluttered videos at scale

Objects: ImageNet Video: T-CNN

Object
Detection

Object
Tracking



(Slides by Andrea Ferri): Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, and Wanli Ouyang, ["Object Detection From Video Tubelets With Convolutional Neural Networks"](#), CVPR 2016 [code]

Objects: Tracking: MDNet

Domain-specific layers are used during training for each sequence, but are replaced by a single one at test time.

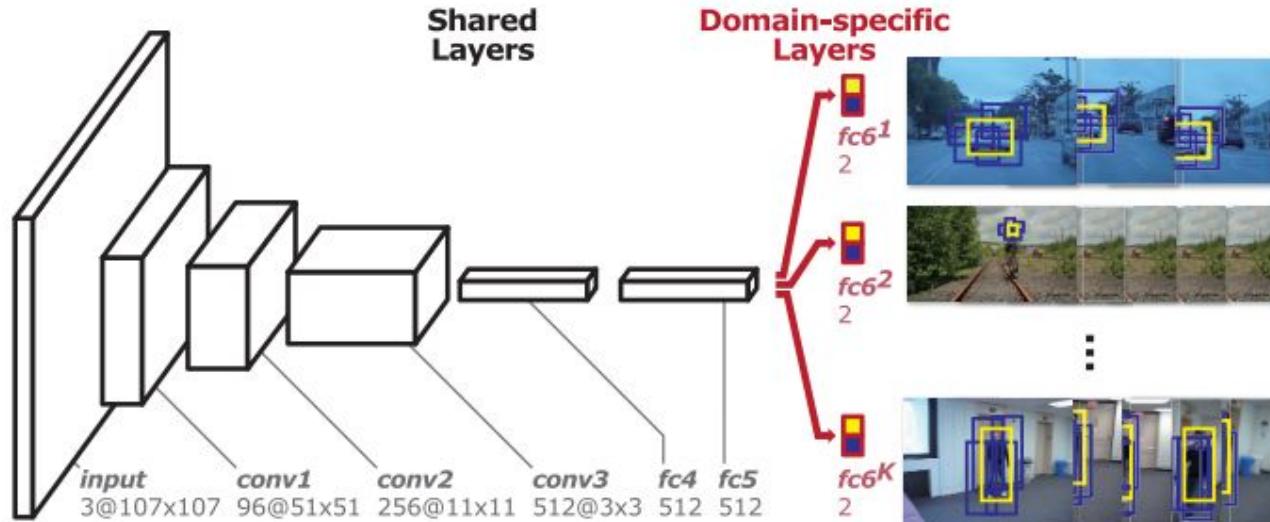
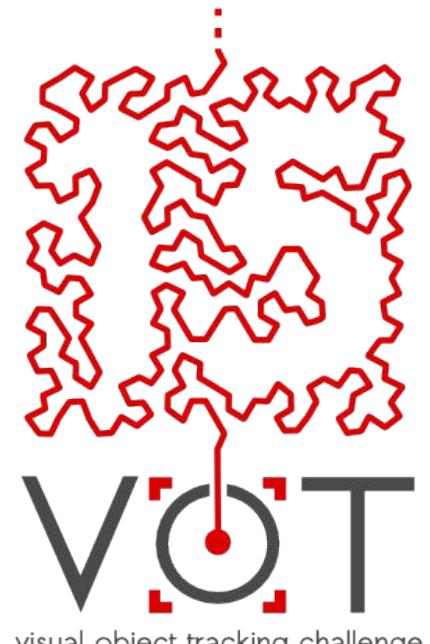


Figure 1: The architecture of our Multi-Domain Network, which consists of shared layers and K branches of domain-specific layers. Yellow and blue bounding boxes denote the positive and negative samples in each domain, respectively.

Nam, Hyeonseob, and Bohyung Han. ["Learning multi-domain convolutional neural networks for visual tracking."](#) ICCV VOT Workshop (2015)

Objects: Tracking: MDNet



Nam, Hyeonseob, and Bohyung Han. "[Learning multi-domain convolutional neural networks for visual tracking.](#)" *ICCV VOT Workshop* (2015)

Objects: Tracking: FCNT

Focus on conv4-3 and conv5-3 of VGG-16 network pre-trained for ImageNet image classification.

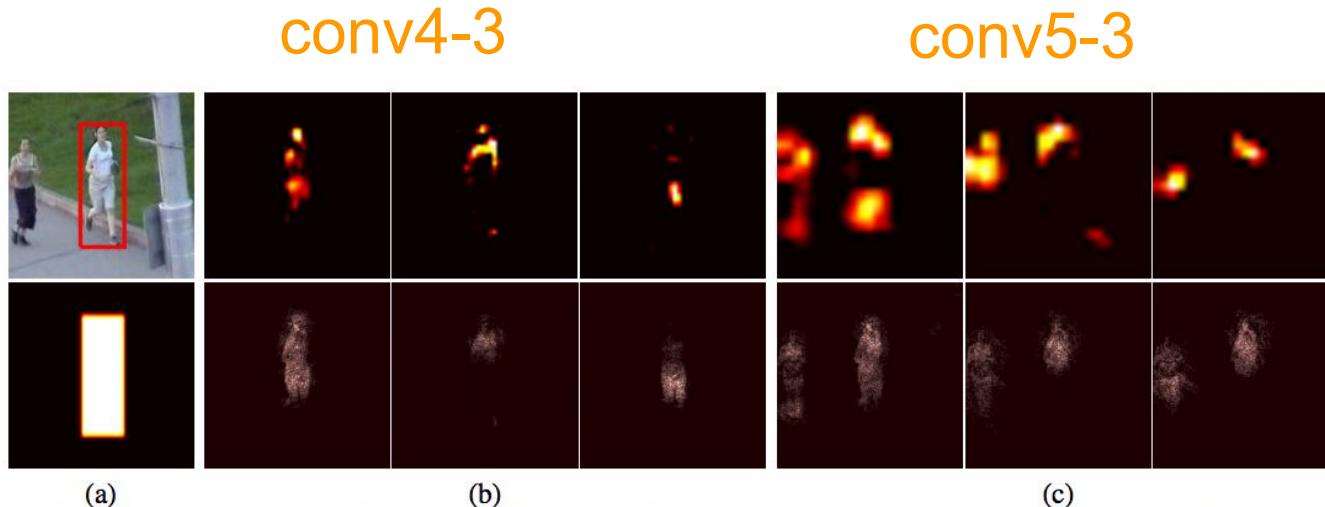
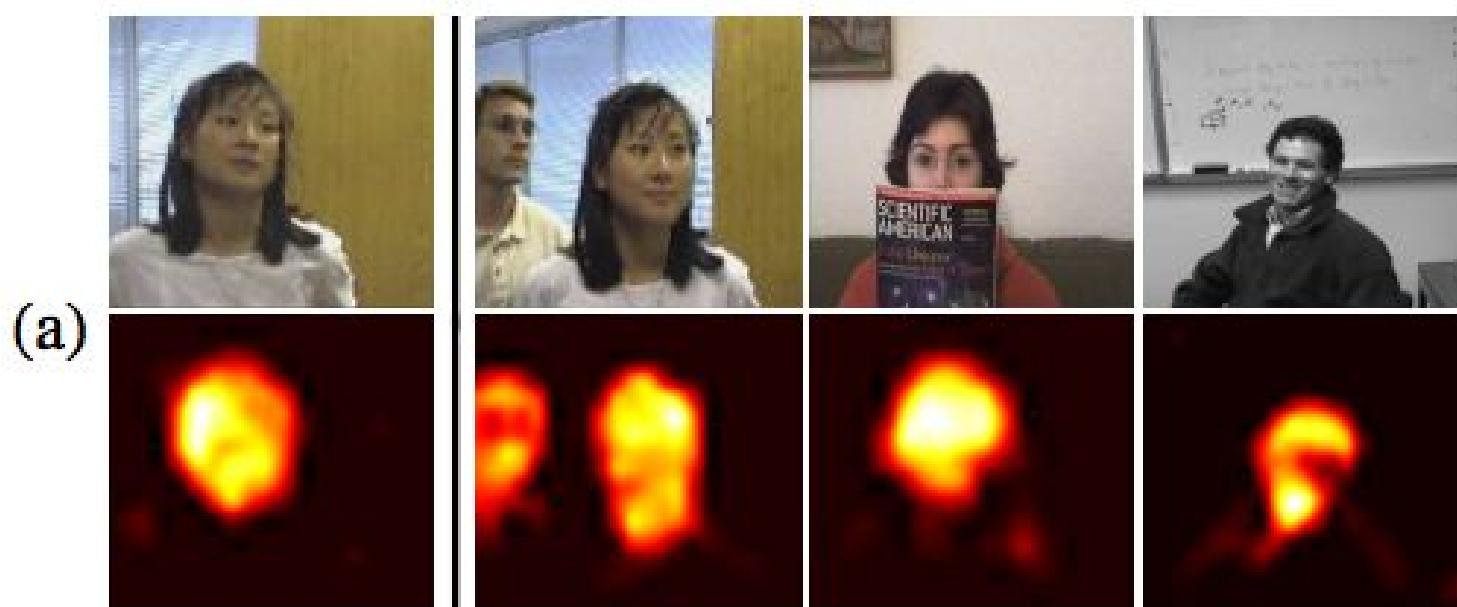


Figure 2. CNNs trained on image classification task carry spatial configuration information. (a) input image (top) and ground truth foreground mask. (b) feature maps (top row) of conv4-3 layer which are activated within the target region and are discriminative to the background distracter. Their associated saliency maps (bottom row) are mainly focused on the target region. (c) feature maps (top row) of conv5-3 layer which are activated within the target region and capture more semantic information of the category (both the target and background distracter). Their saliency maps (bottom row) present spatial information of the category.

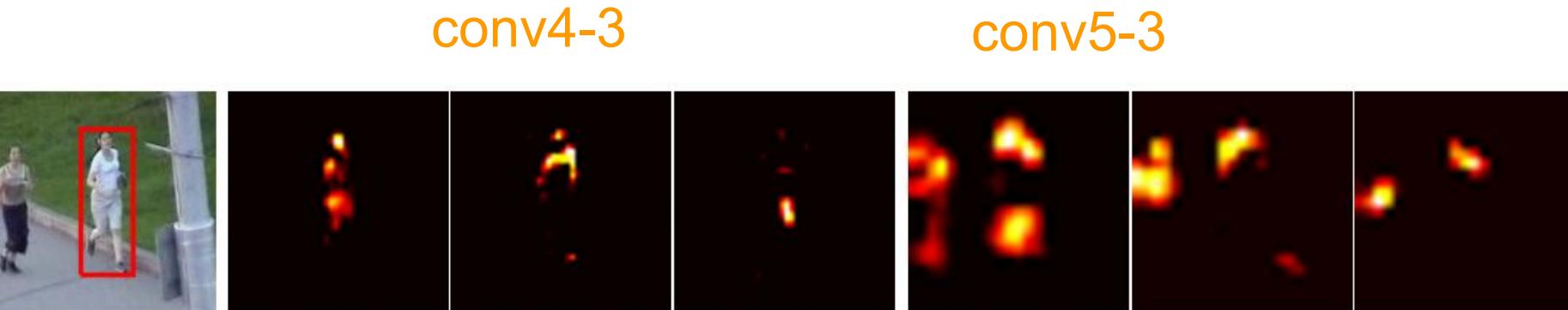
Objects: Tracking: FCNT: Localization

Despite trained for image classification, feature maps in conv5-3 enable object localization...but are not discriminative enough to different instances of the same class.



Objects: Tracking: FCNT: Localization

On the other hand, feature maps from **conv4-3** are more sensitive to intra-class appearance variation...



Objects: Tracking: FCNT: Localization

SNet=Specific Network (online update)

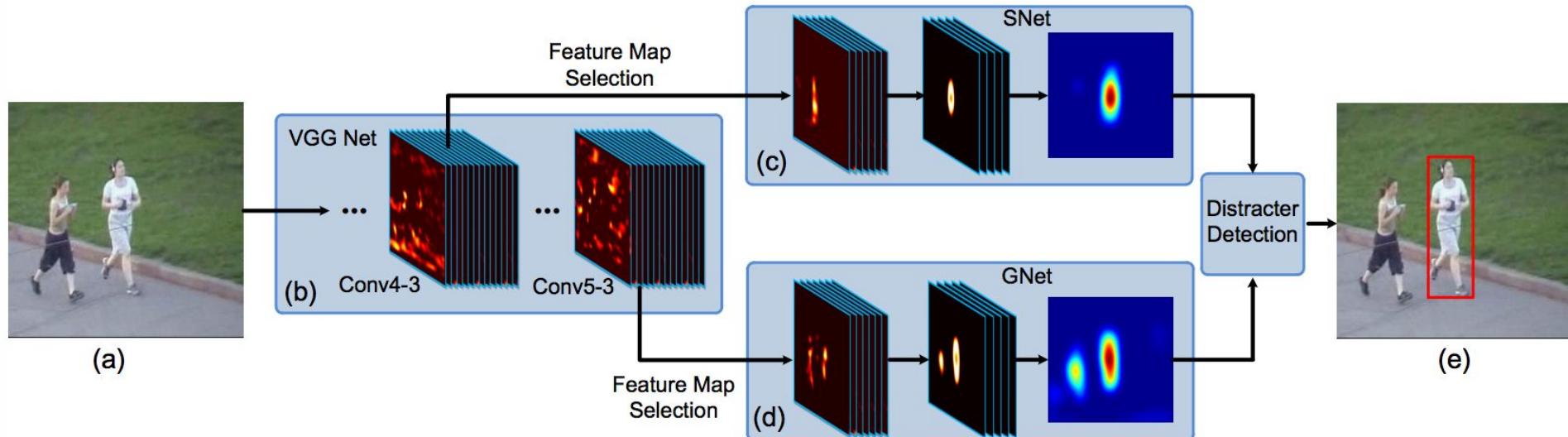


Figure 5. Pipeline of our algorithm. (a) Input ROI region. (b) VGG network. (c) SNet. (d) GNet. (e) Tracking results.

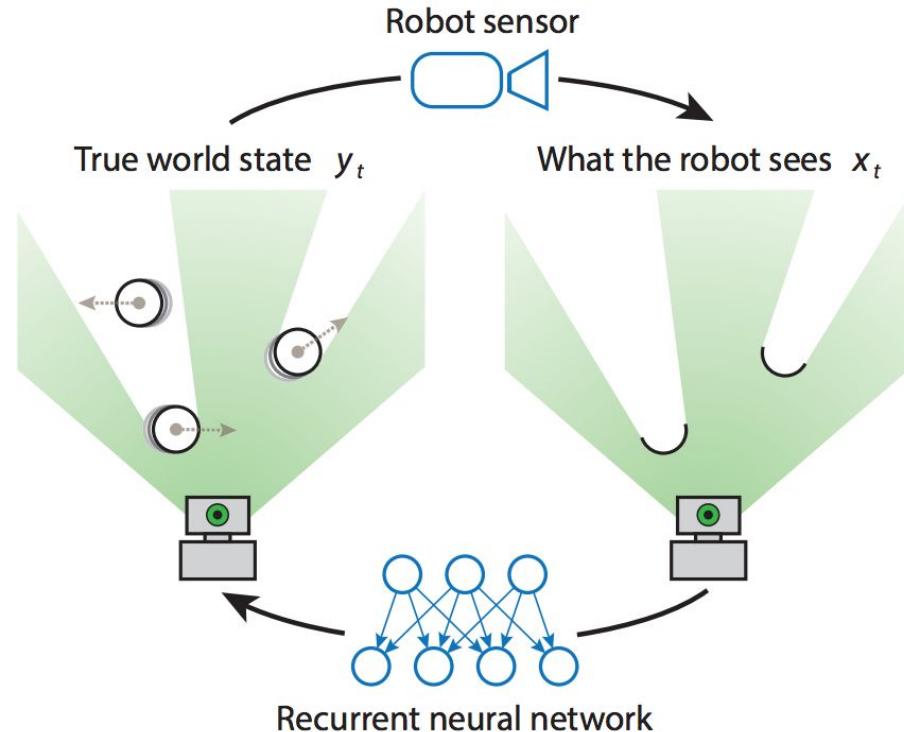
GNet=General Network (fixed)

Objects: Tracking: FCNT: Localization

Other works have also highlighted how features maps in convolutional layers allow object localization.

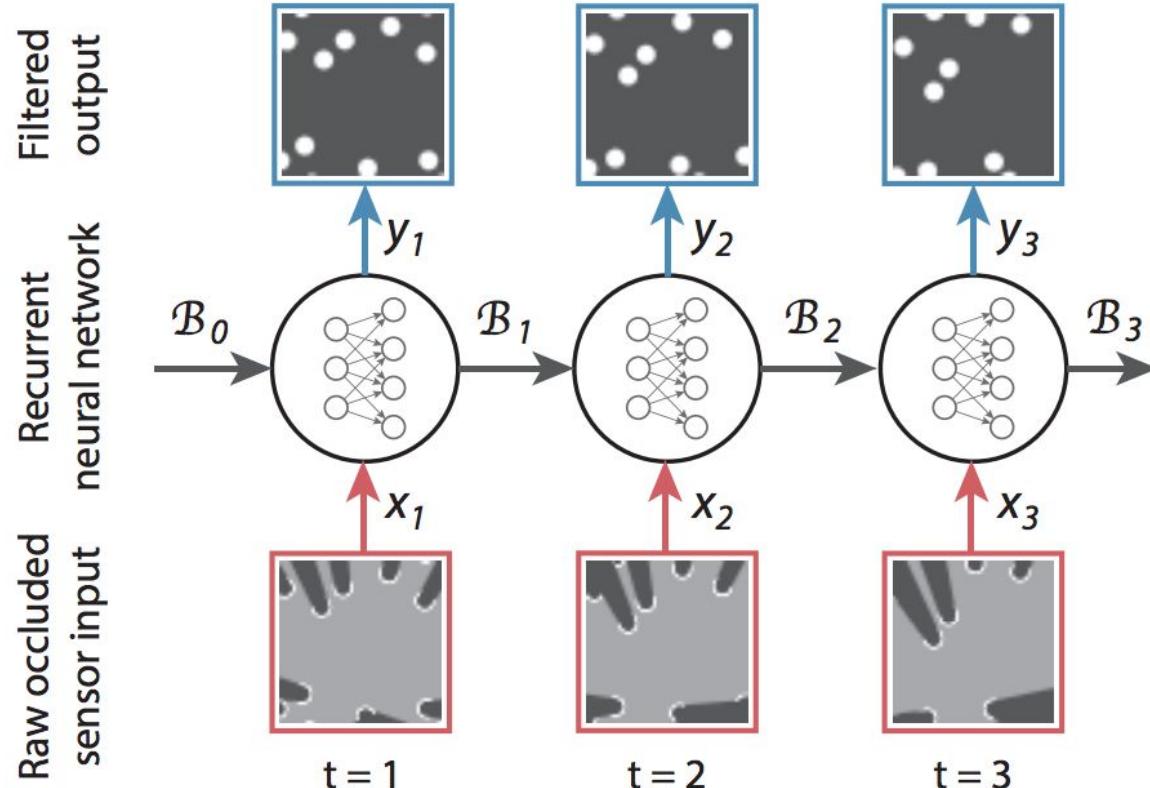


Objects: Tracking: DeepTracking



P. Ondruska and I. Posner, "[Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks](#)," AAAI 2016. [\[code\]](#)

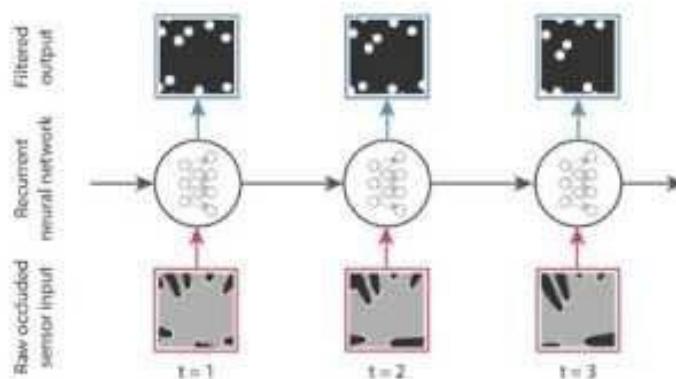
Objects: Tracking: DeepTracking



Objects: Tracking: DeepTracking



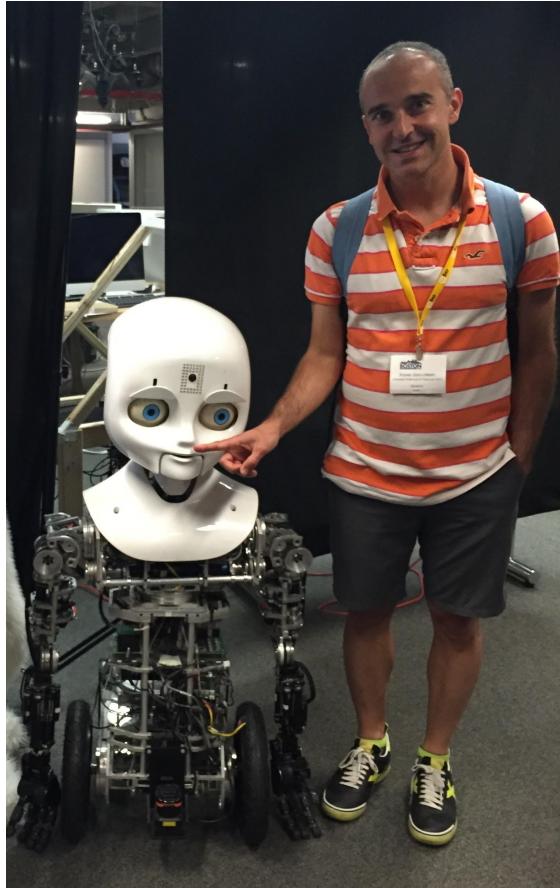
Overview



Summary

- Works on video are normally extensions from principles previously tested on still images.
- RNNs can naturally handle the diversity in video lengths, and capture its temporal dependencies.
- Trick: Init your networks to predict the next frame.

Thanks ! Q&A ?



Follow me at



[/ProfessorXavi](#)



[@DocXavi](#)



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Department of Signal Theory
and Communications

Image Processing Group

<https://imatge.upc.edu/web/people/xavier-giro>