IMAGENET

Large Scale Visual Recognition Challenge 2017 (ILSVRC2017)

Speed/Accuracy Trade-offs for Object Detection from Video

Team Name: IC&USYD

Speaker: Jiankang Deng

Imperial College London

THE UNIVERSITY OF SYDNEY

Submission Brief

- **Object detection from video with provided training data**

Rank 1# mAP: 81.8309%
mAP: 80.8292% (2016 NUIST) **1%**

- **Object detection from video with additional training data**

Rank 1# mAP: 81.9339%

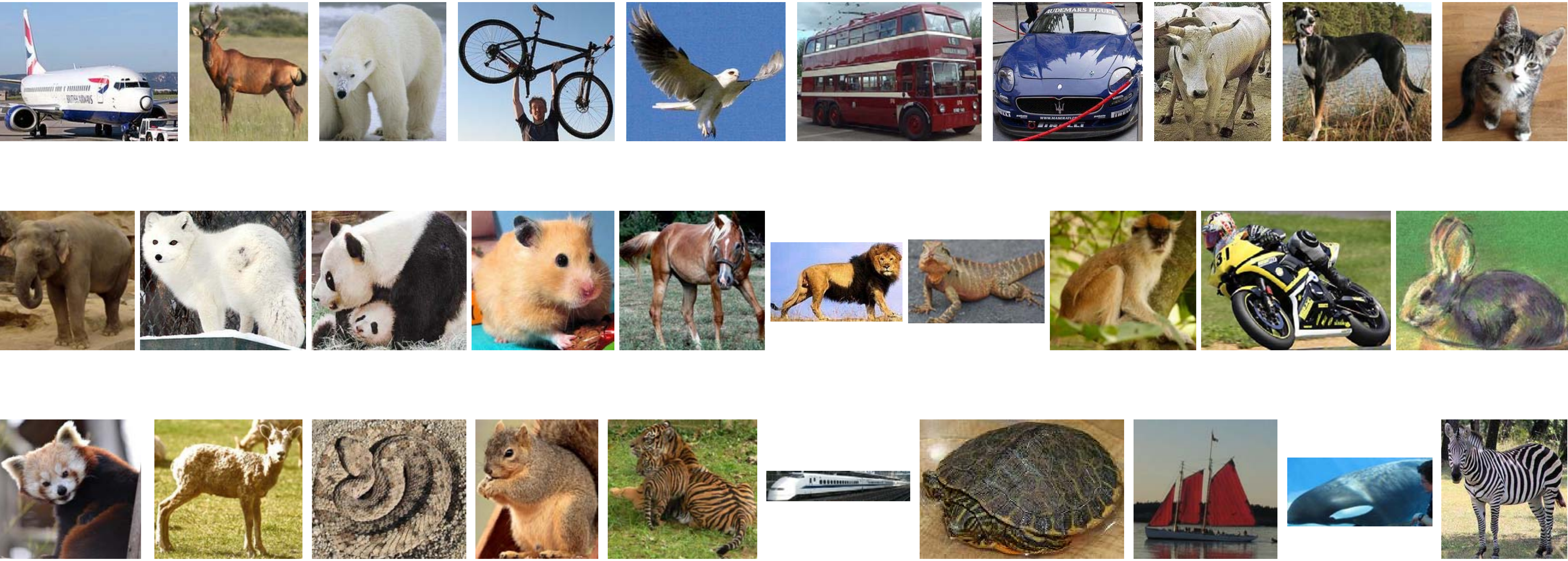
- **Object detection/tracking from video with provided training data**

Rank 1# mAP: 64.1474%
mAP: 55.8557% (2016 CUVideo) **8.29%**

- **Object detection/tracking from video with additional training data**

Rank 1# mAP: 64.2935%

VID Dataset



Class Number: 30

Training set: 4000 snippets
Validation set: 1314 snippets
Test set: 2000 snippets

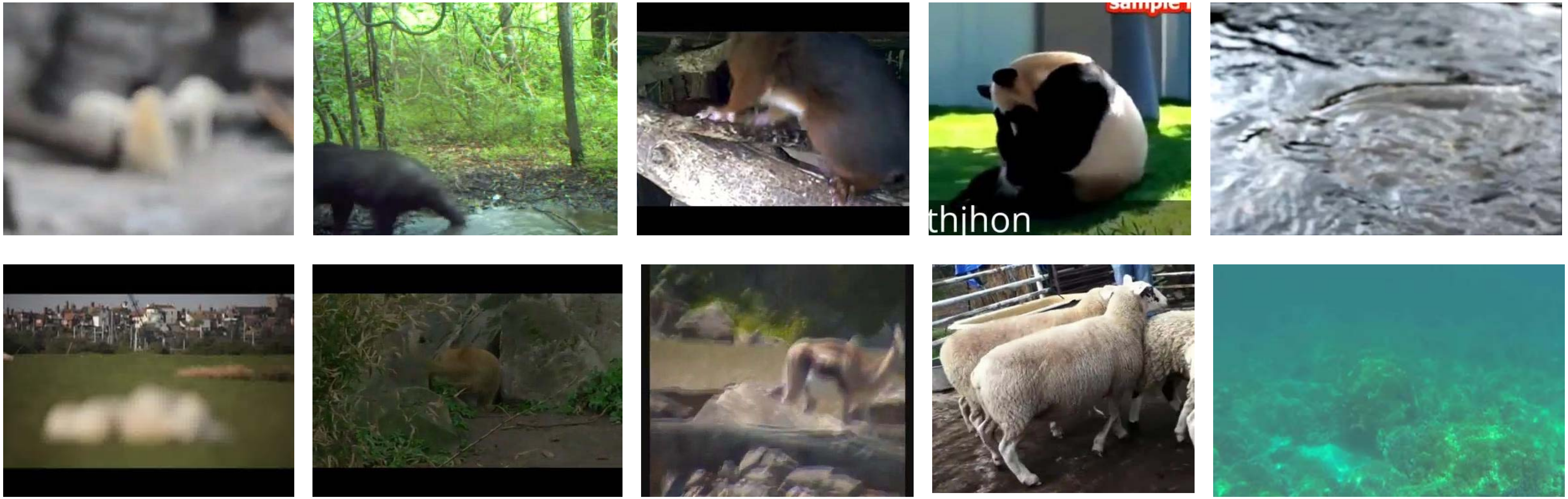
VID Dataset Observation

1. Four leg mammal (18 classes)

2. Vehicle (7 classes)

3. Reptile (3 classes) and context related object, e.g. Bird(sky) and Whale(sea)

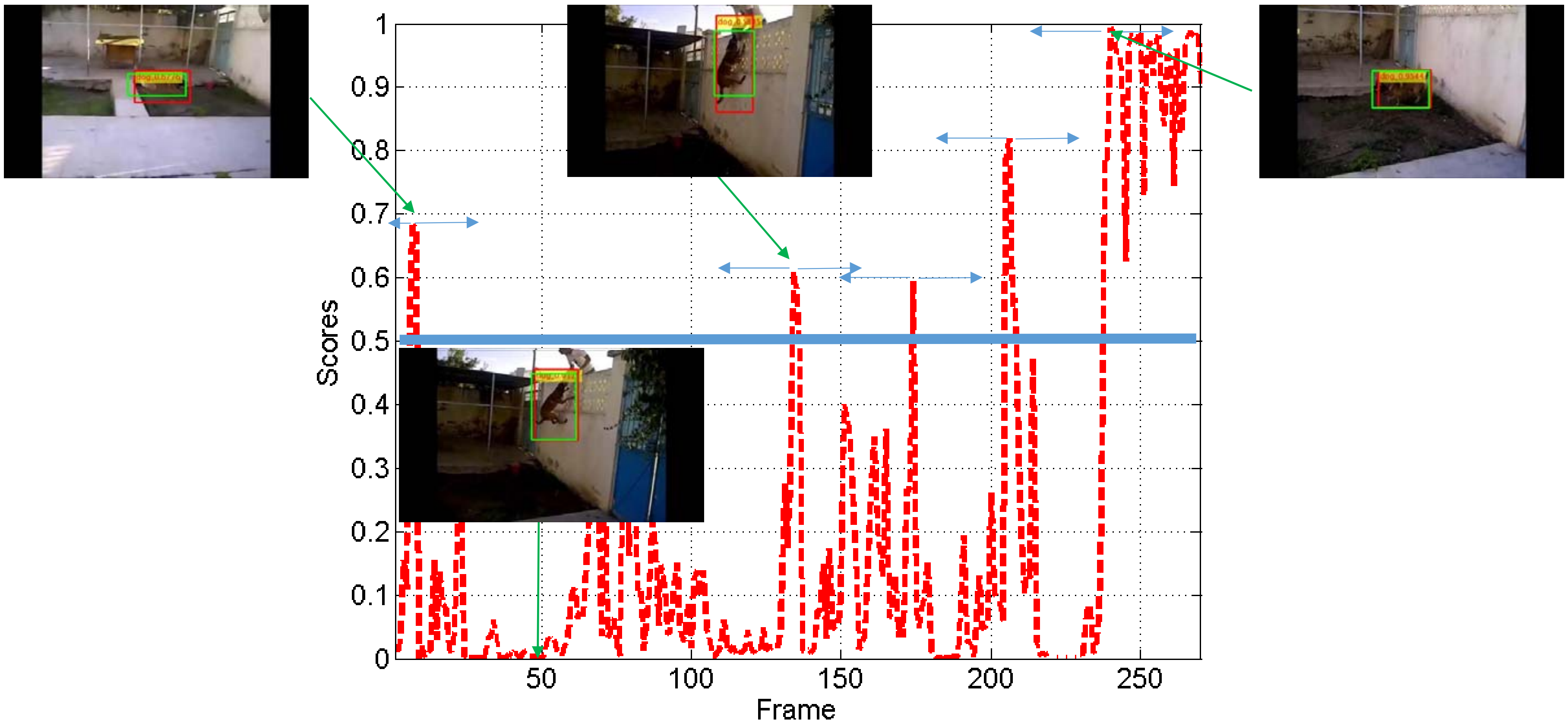
Challenges of VID



camera defocus | **partial occlusion** | **motion blur** | **crowded instance** | **background confusion**

Large appearance variation significantly affects prediction scores.
Temporal information is important to improve the recall.

Submission 2015



- **Object detection on each frame**
- **Object tracking from the high score frames**
- **Box regression and refinement**
- **False Positive suppression by context inference**

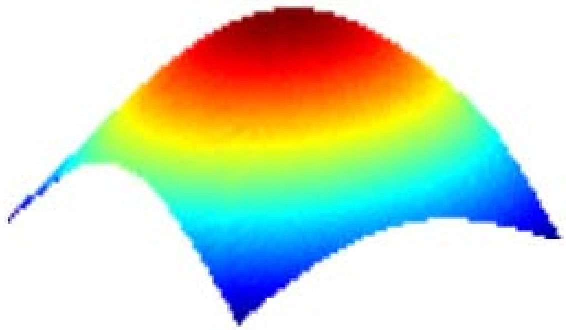
Submission 2016

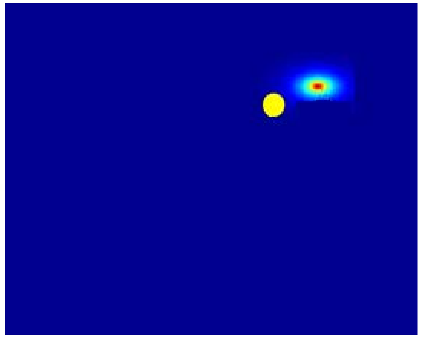
Construct Correlation Filter [1,2] on the Conv Maps

$$r^* = \arg\min_{r} \sum_{i,j}^{W,H} \|r \cdot x_{i,j} - y(i,j)\|_2^2 + \lambda \|r\|_2^2, \qquad y(i,j) = e^{-\frac{(i-W/2)^2+(j-H/2)^2}{2\sigma^2}}$$

FFT

$$R^k = \frac{Y \odot \overline{X}^k}{\sum_{k=1}^{D} X^k \odot \overline{X}^k + \lambda}$$





Correlation Filter Update

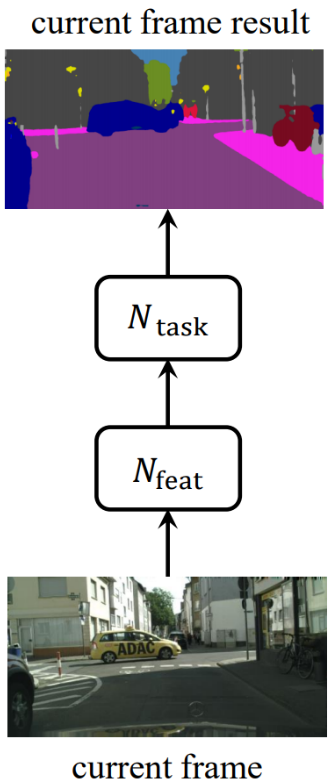
$$A_t^k = \boxed{0.3A_0} + (0.7-\mu)\, A_{t-1}^k + \mu Y \odot \overline{X}_t^k$$
$$B_t^k = \boxed{0.3B_0} + (0.7-\mu)\, B_{t-1}^k + \mu \sum_{k=1}^{D} X_t^k \odot \overline{X}_t^k$$
$$R_t^k = \frac{A_t^k}{B_t^k + \lambda},$$

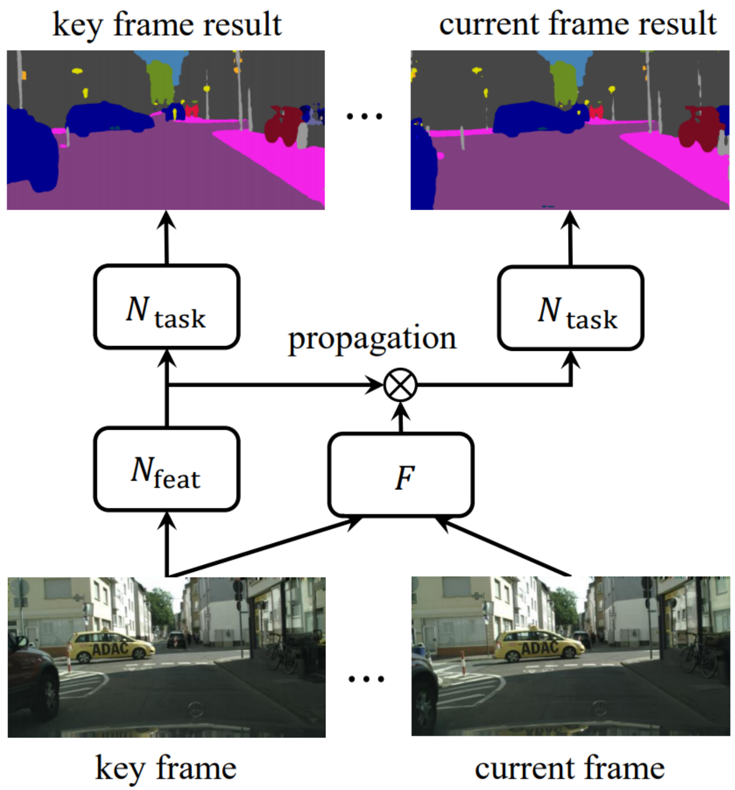
General tracking is object-oriented.
VID task has class-specific prior.

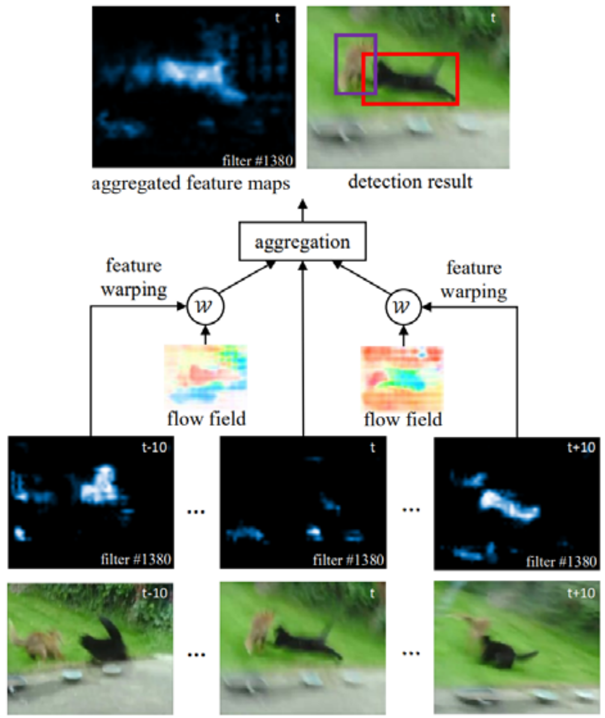
[1] Ma, Chao, et al. "Hierarchical convolutional features for visual tracking." CVPR. 2015.
[2] K. Zhang, Fast Visual Tracking via Dense Spatio-Temporal Context Learning, ECCV 2014

Recent Works on VID

propagate deep features (ResNet101) by flow







Speed: 4.05 fps; mAP: 73.9%

Speed: 20.25fps; mAP: 73.1% [1]

Speed: 1.36fps; mAP: 76.3% [2]

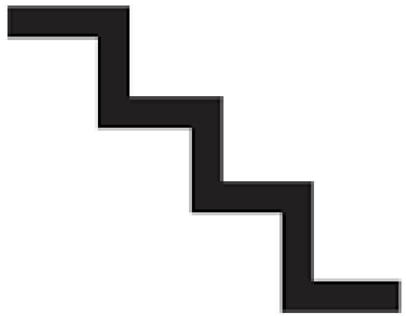
[1]Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep Feature Flow for Video Recognitio.
[2]Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-Guided Feature Aggregation for Video Object Detection.

Training Data and Crowded Status

Training data

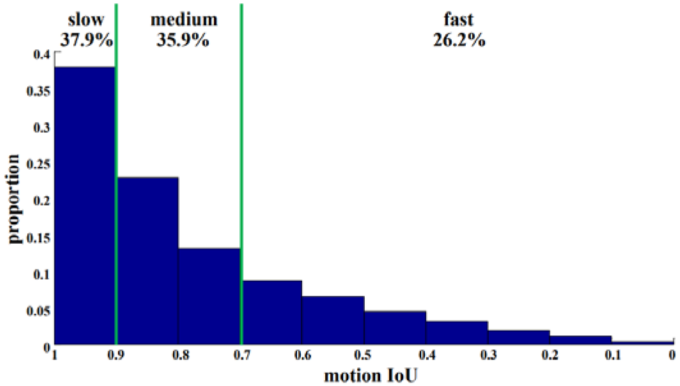
1.Train model on DET data.
2.Predict the score of VID boxes.
3.Select positive examples [0.05, 0.9] from VID.
4. Remove redundant frames (low motion speed).
5. Balance training sample.

Loss: box classification; box regression; crowded status



Adaptive Frame Rate

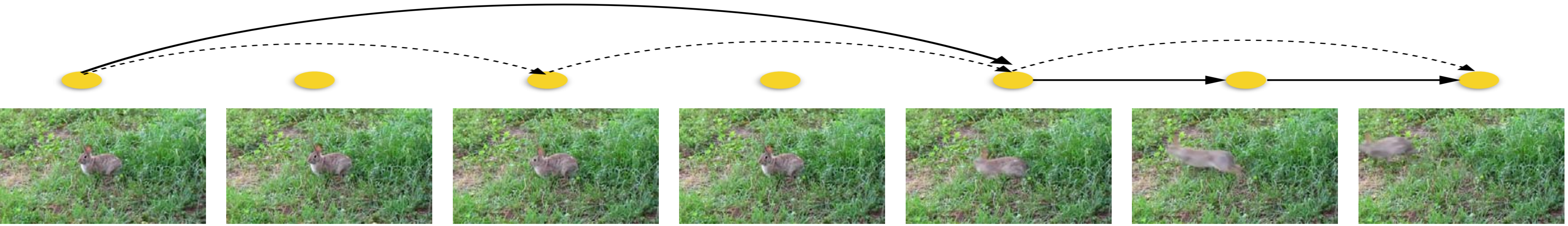
Speed difference



Elephant

Squirrel

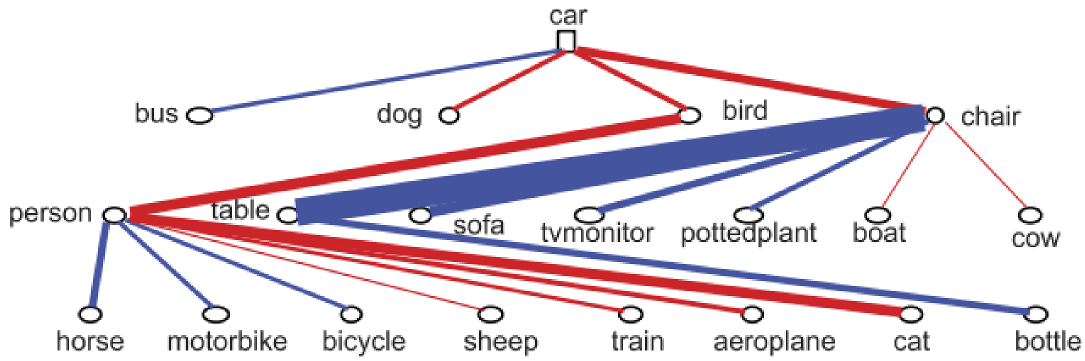
Adaptive frame rate based on motion speed and appearance change



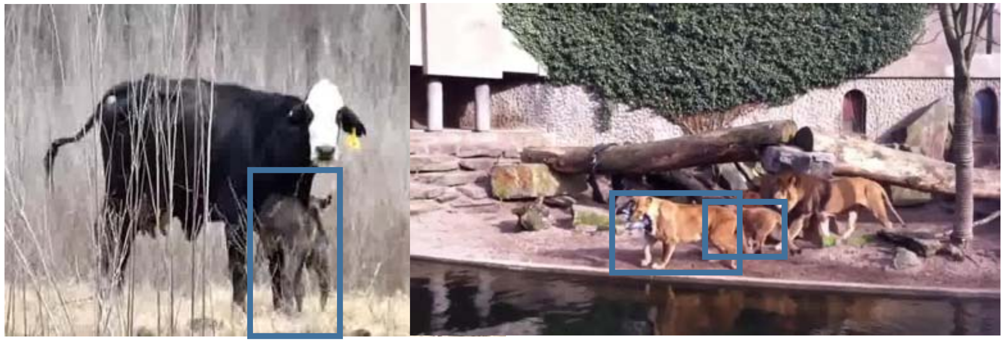
[1] Linchao Zhu, Zhongwen Xu, and Yi Yang. Bidirectional Multirate Reconstruction for Temporal Modeling in Videos.
[2] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos.

Birds of a feather flock together

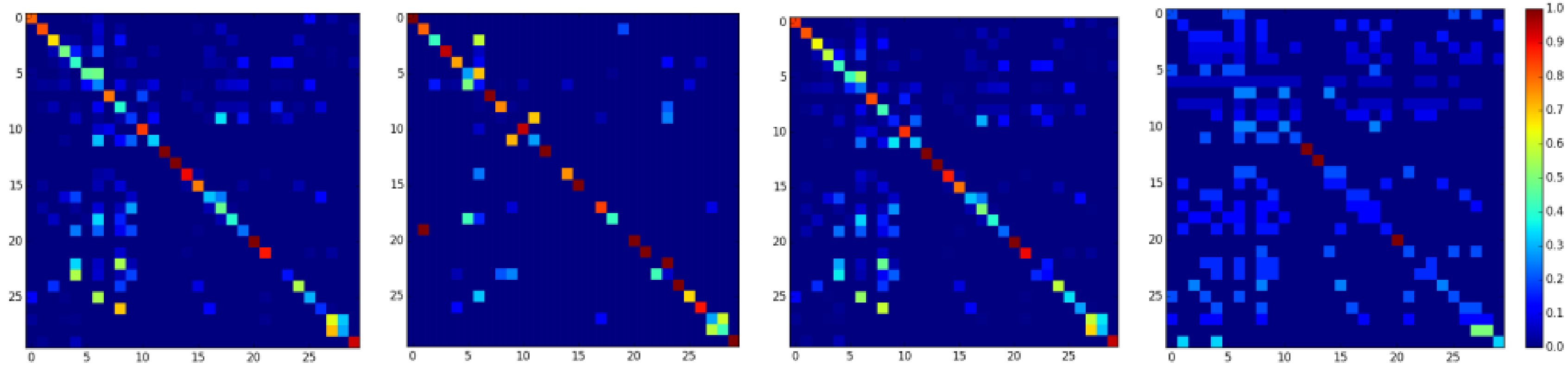
Tree-based context model [1] (VOC 07)



Local appearance is not discriminative.



Co-occurrence matrix on VID

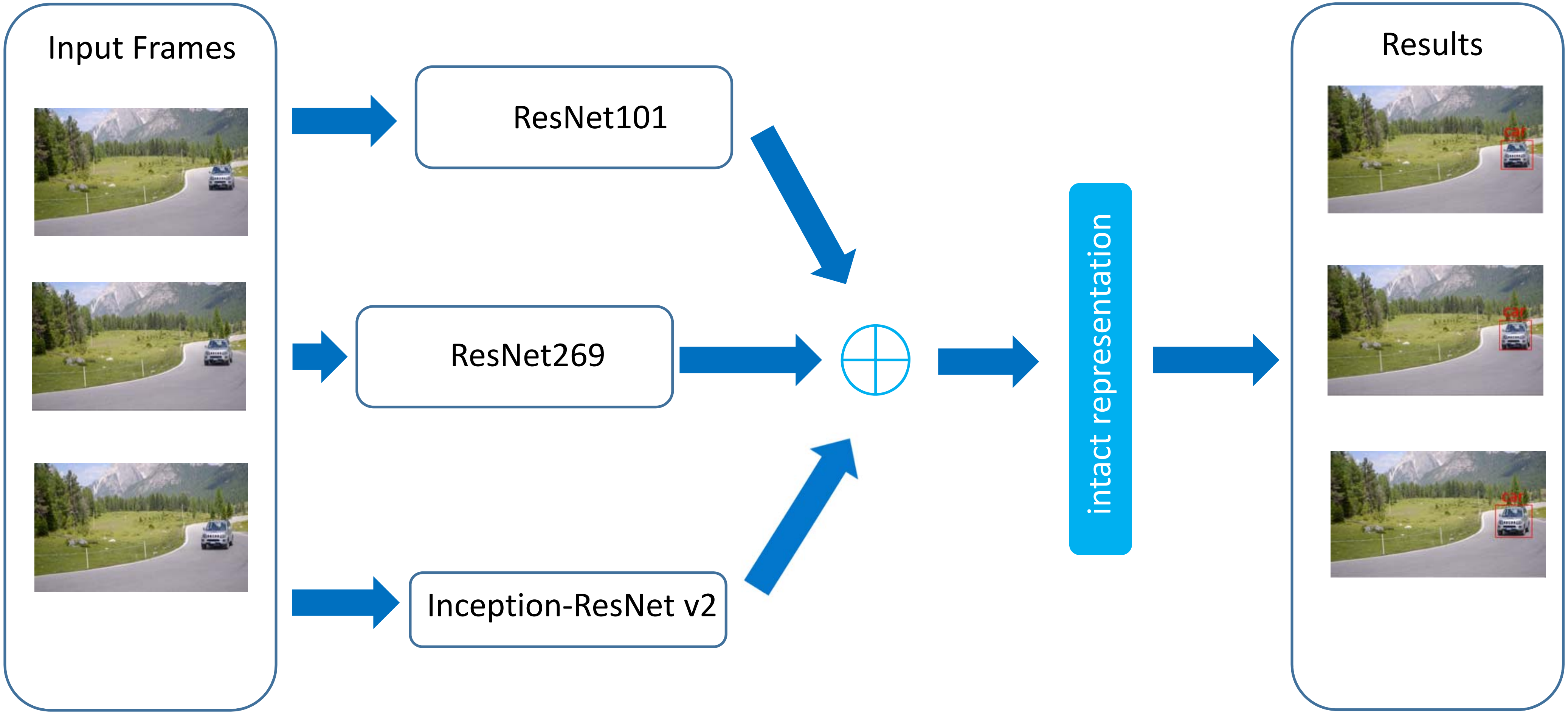


Train **Val** **Train + Val** **Thresholding**

Sparsity 699/900

[1] Choi, Myung Jin. A tree-based context model for object recognition. PAMI, 2012.

Model Ensemble

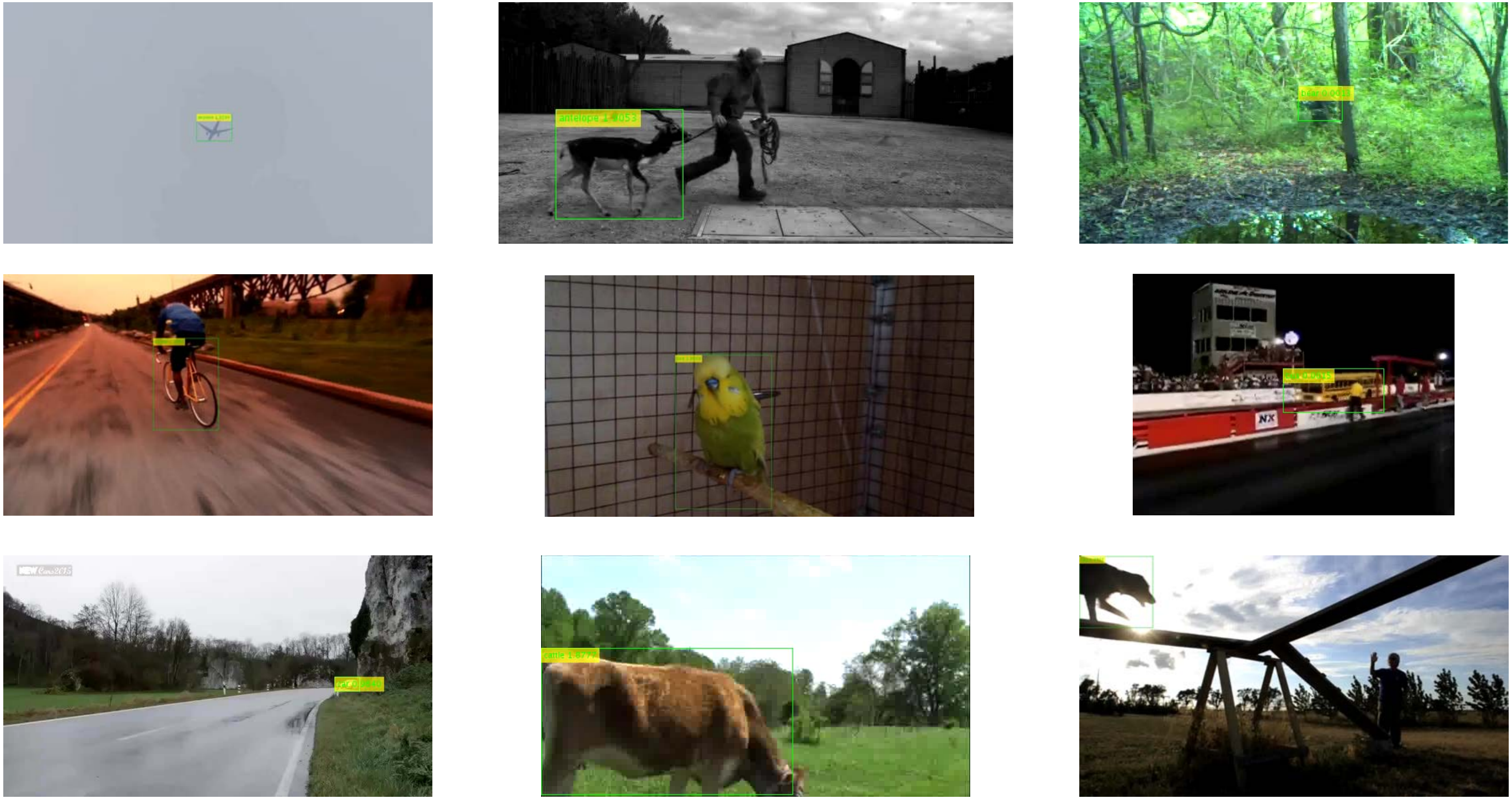


[1] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view Intact Space Learning.

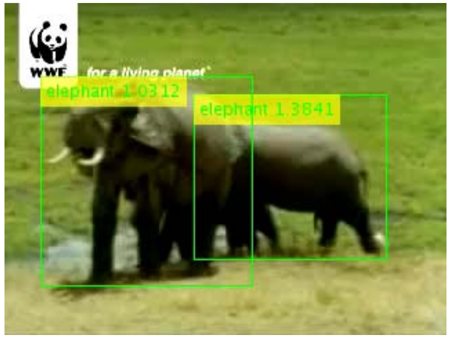
Experimental Results

Method	mAP (%) on the validation set	FPS
Baseline: Single frame R-FCN (ResNet 101)	74.5	4.10
++ Adaptive Frame Rate Deep features propagation and aggregation by flow	**76.8**	**15.4**
++ Context inference (suppress FP)	77.6	
++ Short tractlet combination and re-scoring (similar to seq-NMS)	80.7	
++ Global stage-wise re-rank	82.4	
Submission	mAP (%) on the test set	
++ Ensemble ResNet 269 and Inception-ResNet v2	81.8309	

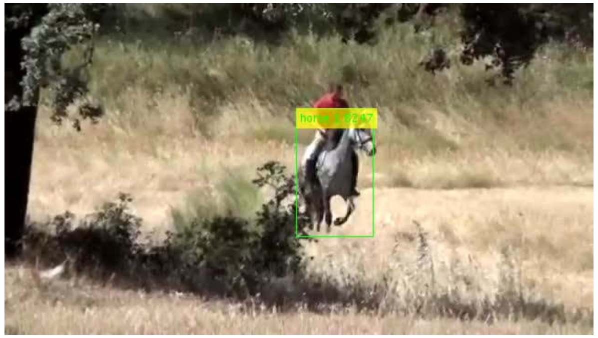
Demo Video

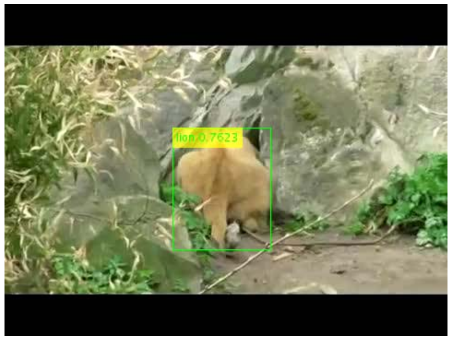


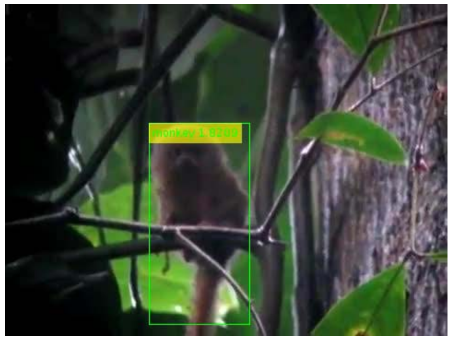
Demo Video


WWF
for a living planet
elephant 1.0312
elephant 1.3841

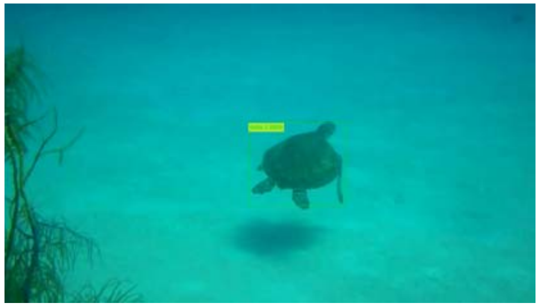

The treats I used over the two weeks included:
flavoured treat drops, vegetables, dog biscuits,
seeds, nuts, mealworms and scrambled egg.

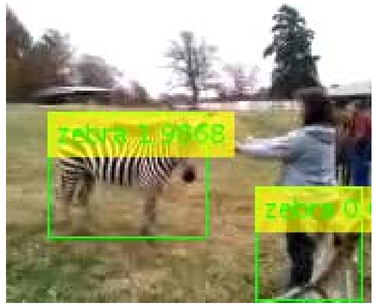



lion 0.7623



Demo Video





Team member






Jiankang Deng[1] Yujiang Zhou[1] Baoshen Yu[2] Zhe Chen[2] Stefanos Zafeiriou[1] Dacheng Tao[2]

1. Intelligent Behavior Understanding Group (ibug), Imperial College London, UK
2. UBTECH Sydney AI Centre, University of Sydney, Australia

Imperial College London

THE UNIVERSITY OF SYDNEY

Email: jiankangdeng@gmail.com