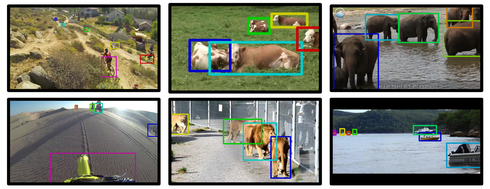
Object Detection from Video Tubelets with Convolutional Neural Networks

Kai Kang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang
Department of Electronic Engineering
The Chinese University of Hong Kong

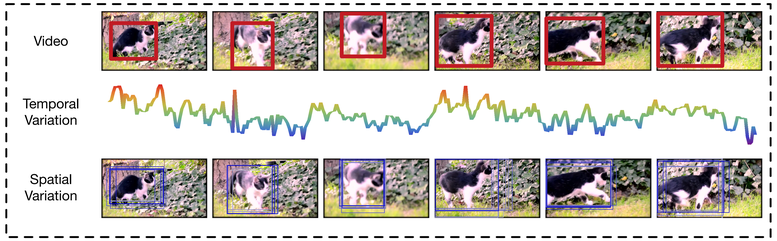
Project Page

Code

Object Detection in Videos

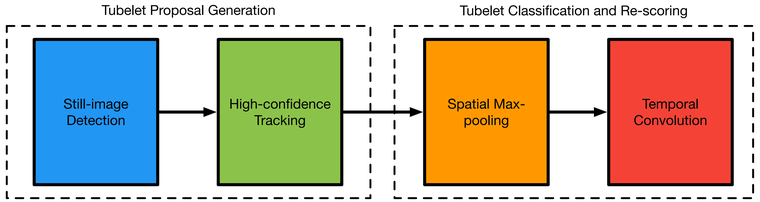


Difficulties of using still-image detectors on videos
- ✓ Large temporal variations
- ✓ Large spatial variations

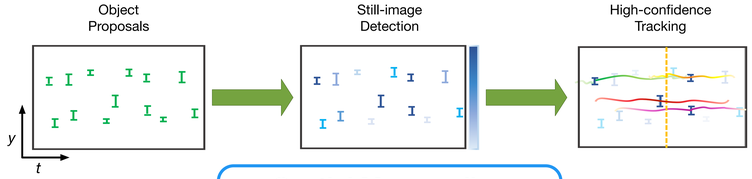


Framework

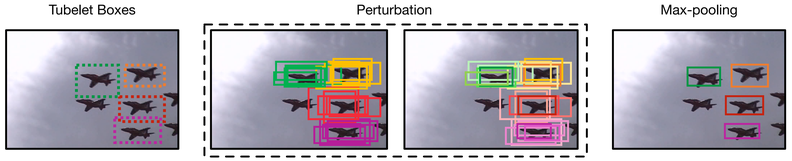
- ✓ A multi-stage framework for object detection in videos
  1. Tubelet proposal generation
  2. Tubelet classification and re-scoring
- ✓ A spatial max-pooling process modeling interactions between detection boxes and tubelets
  - ✓ Largely reduces spatial variations
- ✓ A temporal convolutional network (TCN) to enforce temporal consistency on detection scores
  - ✓ Largely reduces temporal variations



High-confidence Tracking



Spatial Max-pooling



Spatial max-pooling: Adding proposal boxes that have large overlaps with the tubelet boxes and keeping only the boxes with local maximum detection scores. The darker the color, the high the detection scores

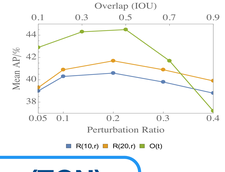
Perturbation schemes:
- ✓Randomly perturbing box coordinates
  - ✓ R(n,r): n boxes with perturbation ratio r
- ✓Adding object proposals (from Selective Search, etc) that have large overlaps with tubelet boxes
  - ✓ O(t): proposals with IOU overlaps of t

Table 1. Performance of sampling perturbation configurations

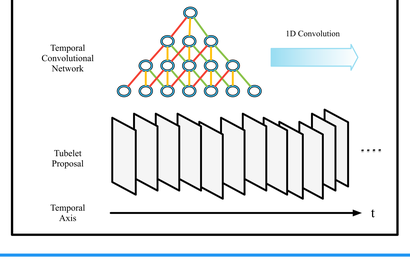
Mean AP / %		Sampling Ratio				
		0.05	0.1	0.2	0.3	0.4
	Tubelet Only			37.4		
# samples	10	39.0	40.3	40.6	39.8	38.8
	20	39.3	40.9	**41.7**	40.9	39.9

Table 2. Performance of overlapping perturbation configurations

Overlap (IOU)	0.1	0.3	0.5	0.7	0.9	Track only
Mean AP / %	42.9	44.3	**44.5**	41.7	37.2	37.4



Temporal Convolution Networks (TCN)



Input Features:
- ✓ Detection scores
- ✓ Tracking confidences
- ✓ Anchor (tracking starting frame) offset distances

Ouputs:
- ✓ Detection probabilities

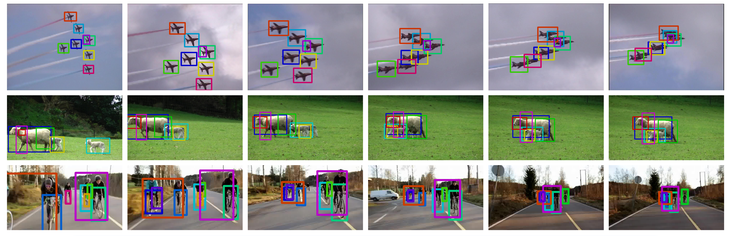
Table 3. TCN setting

	Input	conv1	conv2	conv3	conv4	softmax
Kernel size (n × t)		256 × 5	256 × 5	256 × 7	2 × 3	
Output size (c × t)	3 × 50	256 × 50	256 × 50	256 × 50	2 × 50	2 × 50

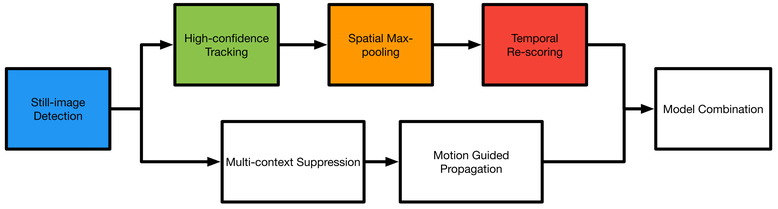
Results on ImageNet VID Dataset

Table 4. Results on ImageNet VID validataion set (inital release)

Method	airplane	antelope	bear	bicycle	bird	bus	car	cattle	dog	cat	elephant	fox	giant panda	hamster	horse	lion
Still-image	64.50	71.40	**42.60**	36.40	18.80	62.40	**37.30**	47.60	15.60	49.50	**66.90**	66.30	58.20	74.10	25.50	29.00
High-confidence Tracking	64.10	61.80	37.60	37.80	19.90	51.10	28.10	46.50	9.50	44.10	52.30	56.40	52.00	56.40	21.20	29.70
Spatial Max-pooling	71.30	70.90	42.50	37.80	21.00	61.70	35.00	48.10	16.60	**53.20**	63.00	70.00	58.00	70.90	27.00	33.50
Temporal convolution	**72.70**	**75.50**	42.20	**39.50**	**25.00**	**64.10**	36.30	**51.10**	**24.40**	48.60	65.60	**73.90**	**61.70**	**82.40**	**30.80**	**34.40**
Method	lizard	monkey	motorcycle	rabbit	red panda	sheep	snake	squirrel	tiger	train	turtle	watercraft	whole	zebra		Mean AP
Still-image	**68.70**	**1.90**	50.80	34.20	**29.40**	**59.00**	**43.70**	1.80	33.00	**56.60**	66.10	61.10	24.10	64.20		45.30
High-confidence Tracking	36.70	1.50	41.40	34.10	19.60	45.50	10.90	1.30	12.80	48.10	64.90	52.40	17.20	63.60		37.40
Spatial Max-pooling	64.30	**1.90**	51.40	34.10	15.50	57.90	40.20	2.40	31.60	52.20	**67.50**	63.10	24.50	67.10		45.10
Temporal convolution	54.20	1.60	**61.00**	**36.60**	19.70	55.00	38.90	**2.60**	**42.80**	54.60	66.10	**69.20**	**26.50**	**68.60**		**47.50**



Part of ImageNet 2015 Winning Framework



Ranked **No. 1** in object detection from video with provided data in ILSVRC 2015 Challenge (CUvideo)

Evaluation on Youtube-Objects Dataset

Table 5. Localization accuracy on Youtube-Objects Dataset

Method	aero	bird	boat	car	cat	cow	dog	horse	mbike	train	Avg.
Prest et al. (CVPR 2012)	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25	28.5
Joulin et al. (ECCV 2014)	25.1	31.2	27.8	38.5	41.2	28.4	33.9	35.6	23.1	25.0	31.0
Kwak et al. (ICCV 2015)	56.5	66.4	58.0	76.8	39.9	**69.3**	50.4	56.3	53.0	31.0	55.7
Ours	**94.1**	**69.7**	**88.2**	**79.3**	**76.6**	18.6	**89.6**	**89.0**	**87.3**	**75.3**	**76.8**