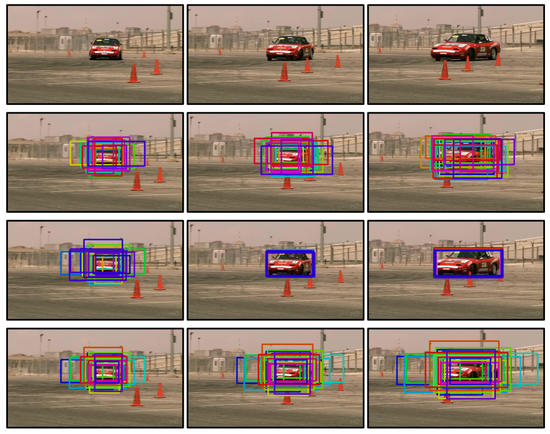
Object Detection in Videos with Tubelet Proposal Networks

Kai Kang[1], Hongsheng Li[1], Tong Xiao[1], Wanli Ouyang[1,4], Junjie Yan[2], Xihui Liu[3], Xiaogang Wang[1]
[1]The Chinese University of Hong Kong, [2]SenseTime Group Limited
[3]Tsinghua University, [4]The University of Sydney

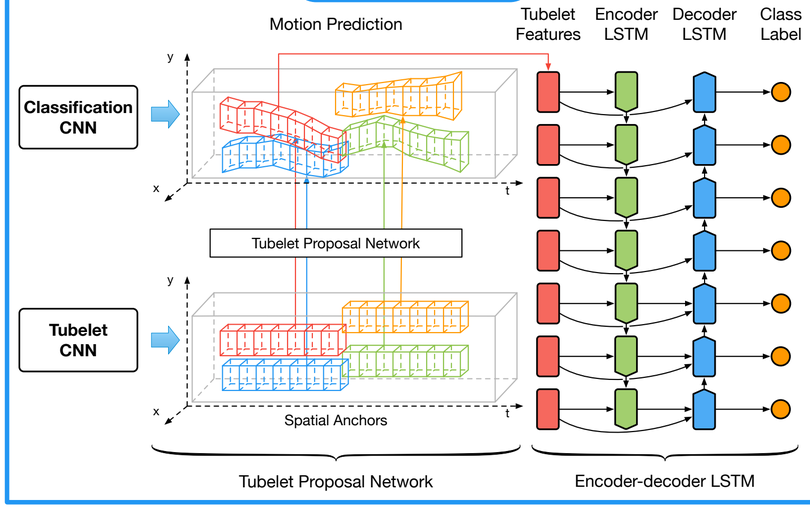
CVPR July 21-26 HONOLULU 2017

Proposals in Video Object Detection

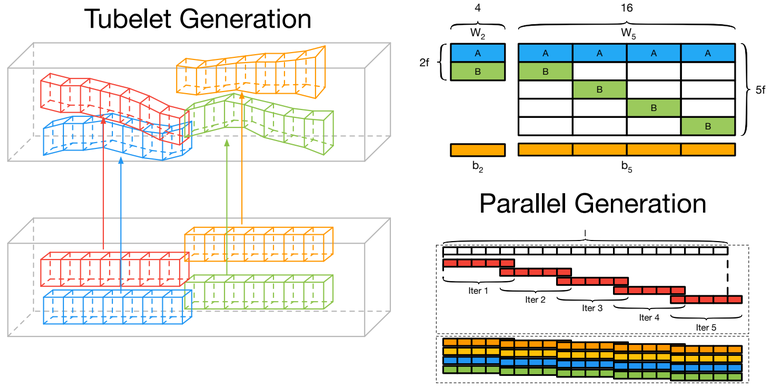


Proposal Method	Box Location Variety	Box Location Stability	Temporal Correspondence
Static proposals	☑	☐	☐
Location Prediction	☐	☑	☑
Motion Prediction	☑	☑	☑

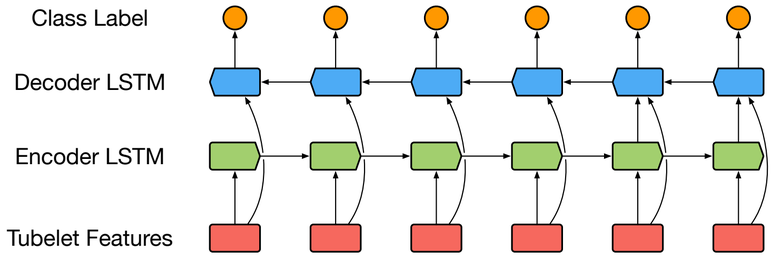
Framework



Tubelet Proposal Networks



Encoder-decoder LSTM



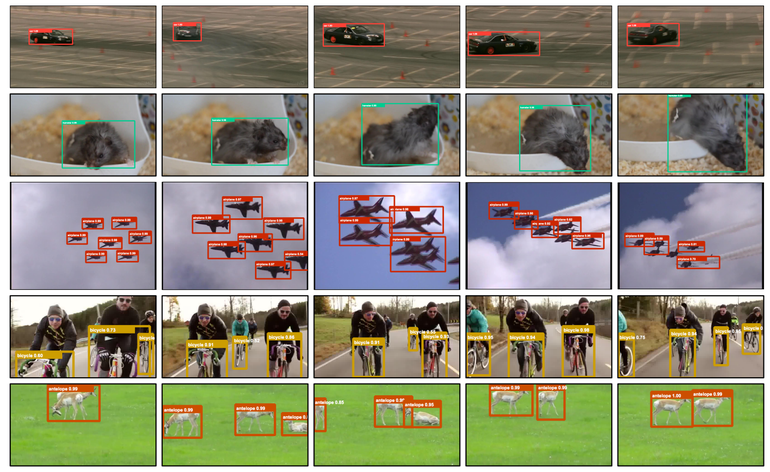
Experimental Settings

Initialization of tubelet proposal networks

Method	Window	Mean Absolute Difference	Mean Relative Difference	Mean IOU
RegressMove Random	2	15.50	0.0730	0.7966
RegressMove Random	5	26.00	0.1319	0.6972
RegressMove RNN	5	13.87	0.0683	0.8060
RegressMove Block	5	**12.98**	**0.0616**	**0.8244**
RegressMove Block	11	15.20	0.0761	0.8017
RegressMove Block	20	18.03	0.0874	0.7731

Qualitative Results

Qualitative results on ImageNet VID validation set



Results on ImageNet VID Dataset

Results on ImageNet VID validation set

Method	airplane	antelope	bear	bicycle	bird	bus	car	cattle	dog	cat	elephant	fox	giant panda	hamster	horse	lion
Static (Fast RCNN)	82.10	78.40	66.50	65.60	66.10	77.20	52.30	49.10	57.10	72.00	68.10	76.80	71.80	89.70	65.10	20.10
RegressMove+Fast RCNN	77.60	77.80	66.30	65.40	64.90	76.60	51.40	49.30	55.90	72.40	68.40	77.50	71.00	90.00	64.20	20.80
RegressBox+LSTM	75.90	78.30	66.00	64.60	68.20	81.30	53.80	52.80	60.50	72.20	69.80	78.20	72.40	90.10	66.40	21.20
RegressMove+LSTM	83.90	79.40	71.50	65.20	68.30	79.40	53.30	61.50	60.80	76.50	70.50	83.90	76.90	91.60	66.10	15.80
RegressMove+ED-LSTM	84.60	78.10	72.00	67.20	68.00	80.10	54.70	61.20	61.60	78.90	71.60	83.20	78.10	91.50	66.80	21.60
Method	**lizard**	**monkey**	**motorcyl**	**rabbit**	**red panda**	**sheep**	**snake**	**squirrel**	**tiger**	**train**	**turtle**	**watercraft**	**whale**	**zebra**		**Mean AP**
Static (Fast RCNN)	63.80	34.70	74.10	45.70	55.80	54.10	57.20	29.80	81.50	72.00	74.40	55.70	43.20	89.40		63.00
RegressMove+Fast RCNN	64.60	32.00	69.10	45.40	58.20	54.00	56.70	28.60	80.60	73.00	73.70	54.30	41.40	88.50		62.30
RegressBox+LSTM	74.30	33.40	72.70	51.30	55.50	61.30	68.80	42.20	81.30	78.10	76.00	60.90	42.90	87.40		65.30
RegressMove+LSTM	74.60	34.70	77.10	52.50	71.00	60.90	63.70	40.60	84.50	78.60	77.40	60.20	63.70	89.00		67.80
RegressMove+ED-LSTM	74.40	36.60	76.30	51.40	70.60	64.20	61.20	42.30	84.80	78.10	77.20	61.50	66.90	88.50		**68.40**

YouTube-Objects Dataset

Object localization on YouTube-Objects (YTO) 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
Kang et al. (CVPR 2016)	**94.1**	69.7	88.2	79.3	76.6	18.6	89.6	**89.0**	87.3	75.3	76.8
Ours	91.2	**99.4**	**93.1**	**94.8**	**94.3**	**99.3**	**90.2**	87.8	**89.7**	**84.2**	**92.4**