Hierarchical Convolutional Features for Visual Tracking

Chao Ma SJTU

Jia-Bin Huang UIUC

Xiaokang Yang SJTU

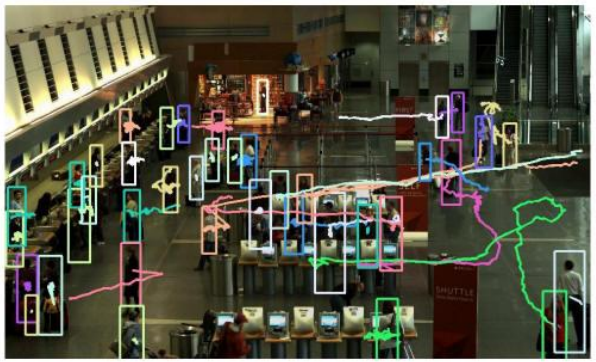
Ming-Husan Yang UC Merced

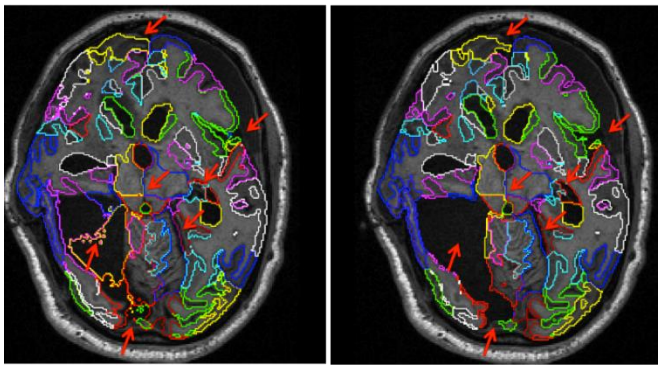
ICCV 2015

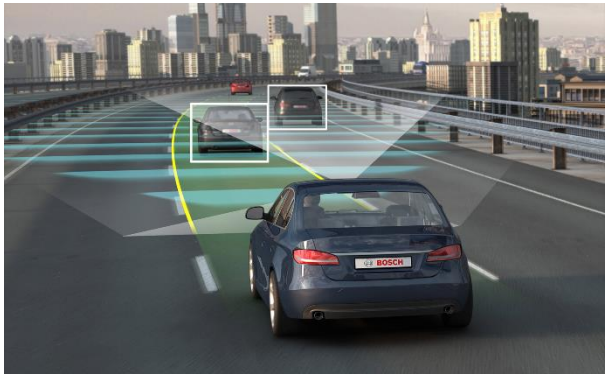
Background

- Given the initial state (position and scale), estimate the unknown states in the subsequence frames
  - Model-free
  - Single target visual tracking

Real-Applications with Tracking





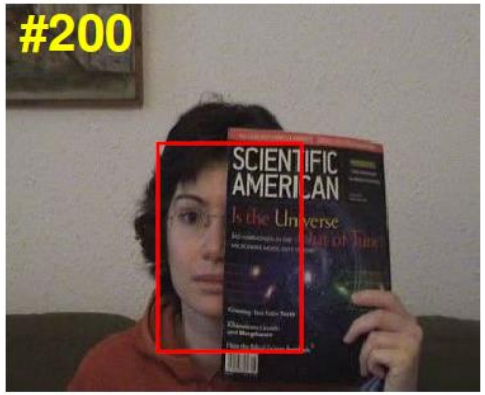


Images from Google Search

Challenges I

(a) Illumination changes

(b) Rotations



(c) Heavy occlusions

(d) Abrupt motions

Challenges II

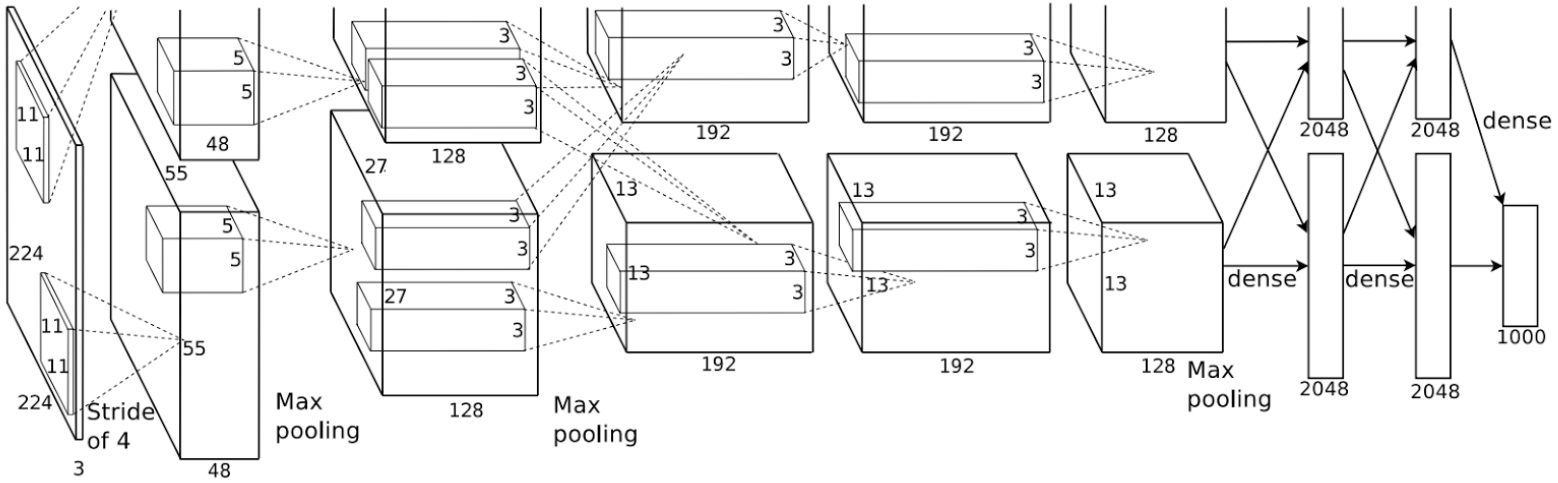
(e) Low resolutions

(f) Scale changes

- Challenges = significant appearance variations over time!!!

Convolutional Neural Networks

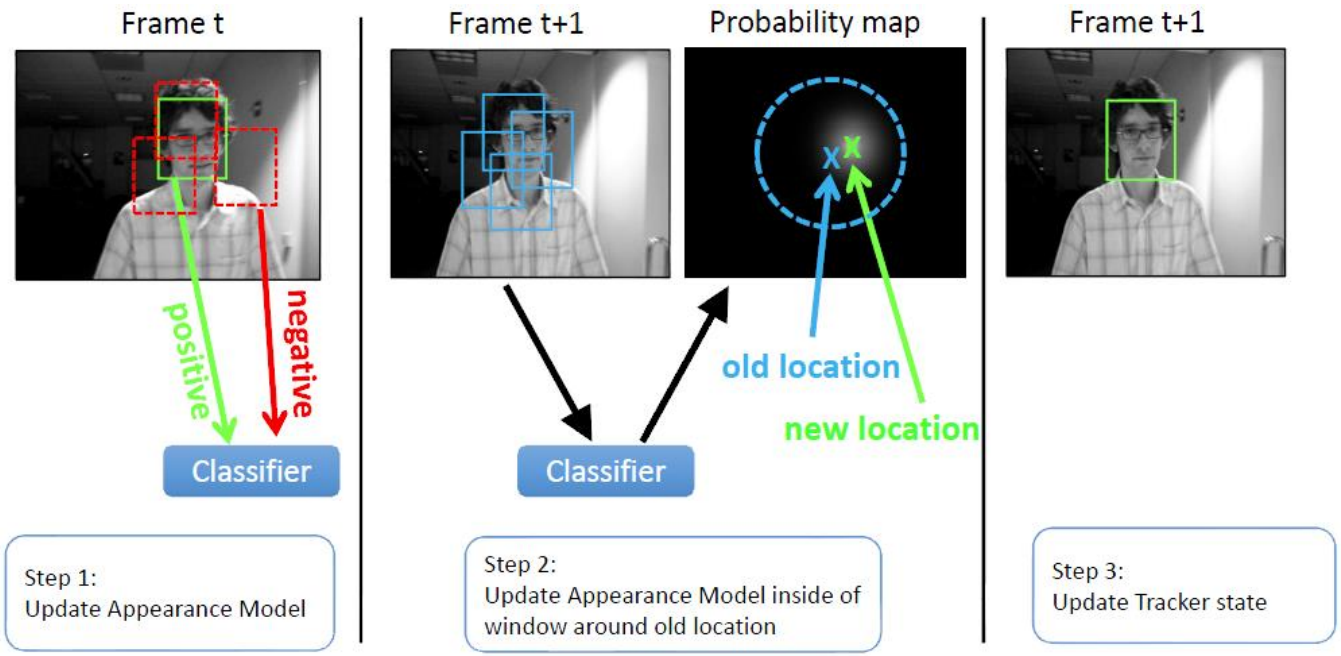
- Show significant advantages on a wide range of computer vision problems: image classification, object detection, object recognition et al.



AlexNet (NIPS'12)

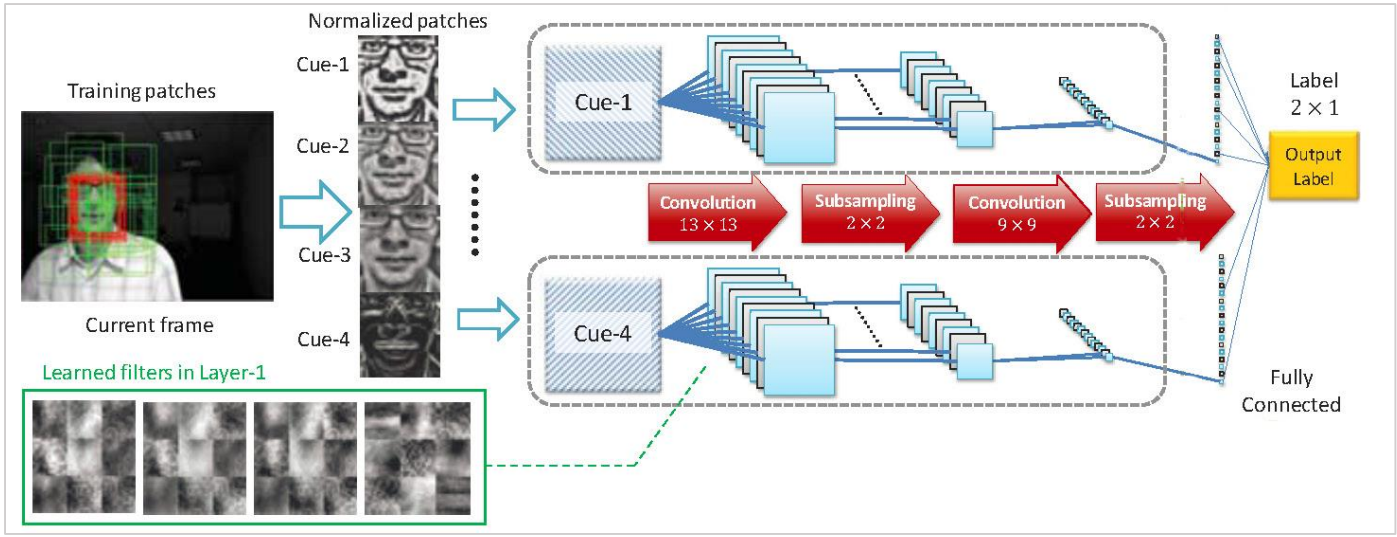
Typical Tracking Framework

- Incrementally learn classifiers to separate targets from background (online learning to adapt to appearance changes)
  - MIL (CVPR'09), Struck (ICCV'11), CT (ECCV'12), ASLA (CVPR'12), MEEM (ECCV'14), etc.



Existing CNN Trackers

- DLT (NIPS'13), LHF (TIP'15), DeepTrack (BMVC'14), CNN-SVM (ICML'15), MDNet (CVPR’16)



This figure credits to Li et al. in the DeepTrack (BMVC’ 14)

Issues of Existing CNN Trackers

- Only use the last (fully-connected) layer of the CNN network for classification
  - Too coarse to localize target precisely
- Sample target states with binary labels (positive and negative)
  - Ambiguity in labeling the spatially over-correlated samples
    - MDNet (CVPR'16): negative mining
    - Struck (ICCV'11): structure output

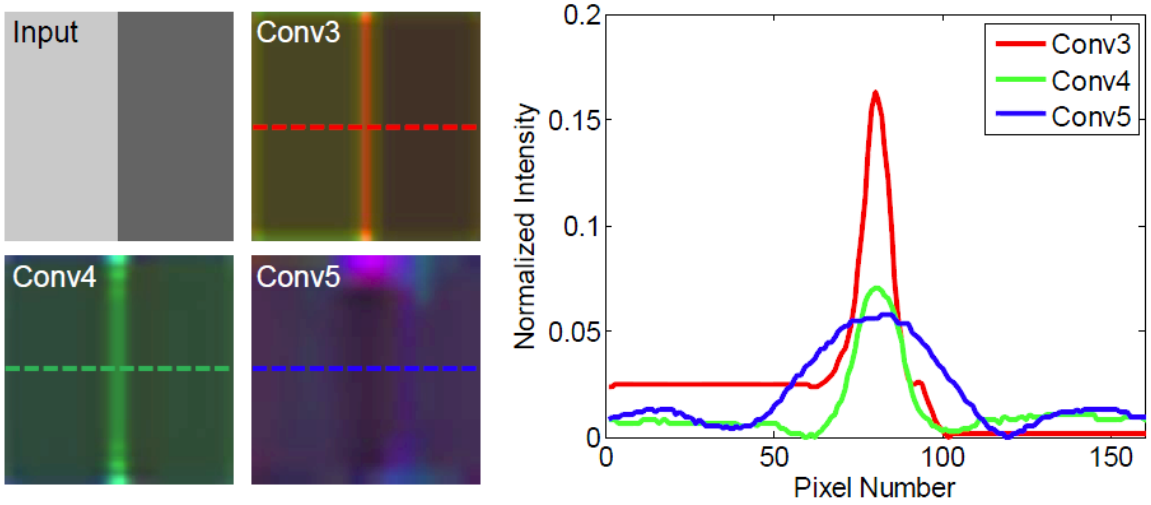
Issues of Existing CNN Trackers

- Only use the last (fully-connected) layer of the CNN network for classification
  - Too coarse to localize target precisely
- Sample target states with binary labels (positive and negative)
  - Ambiguity in labeling the spatially over-correlated samples
    - MDNet (CVPR'16): negative mining
    - Struck (ICCV'11): structure output

Our Observations

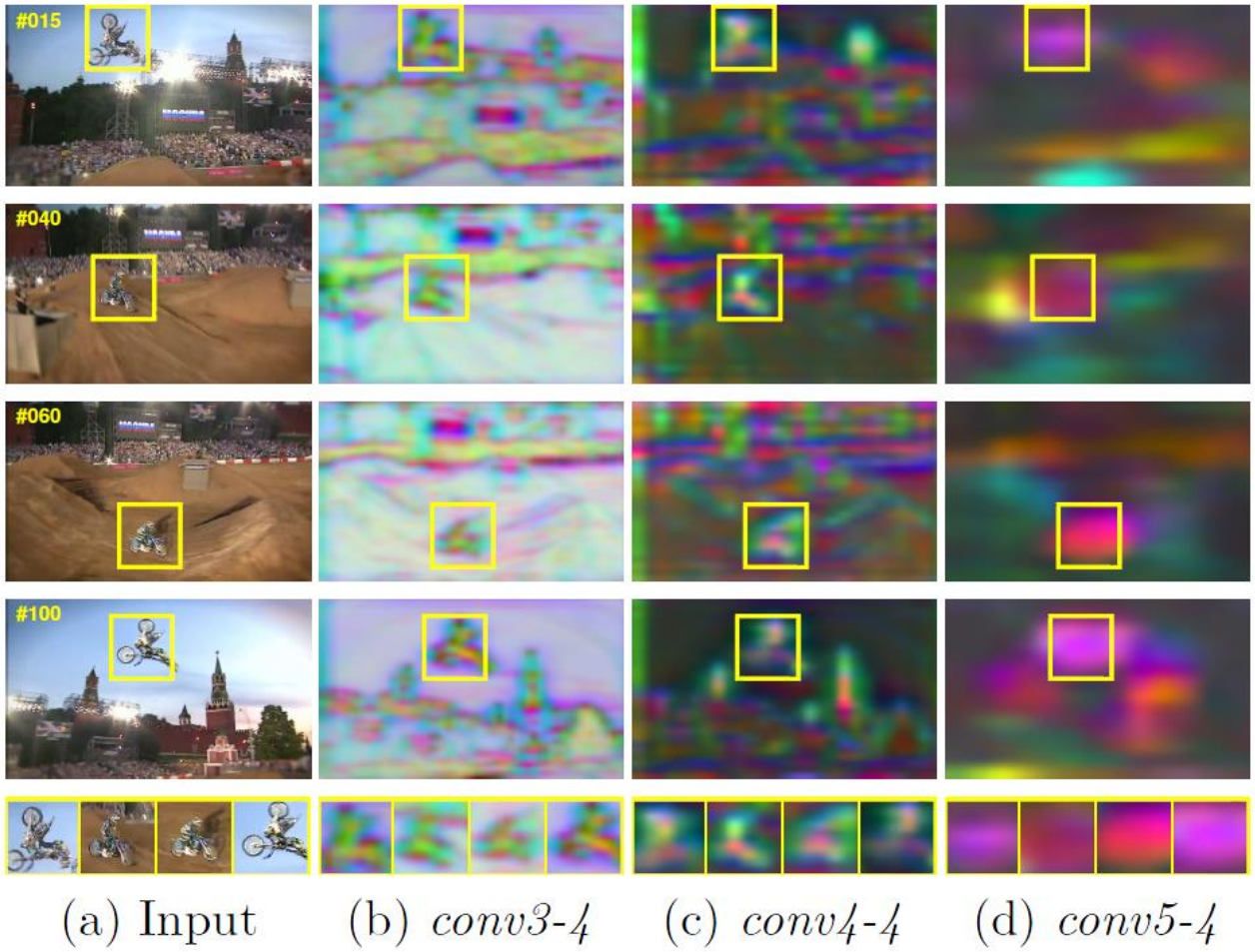
- Earlier layers retain higher spatial resolution for precise localization.
- Latter layers capture more semantic information and are robust to appearance changes.
- Exploit the rich hierarchies for robust visual tracking.

Toy Example

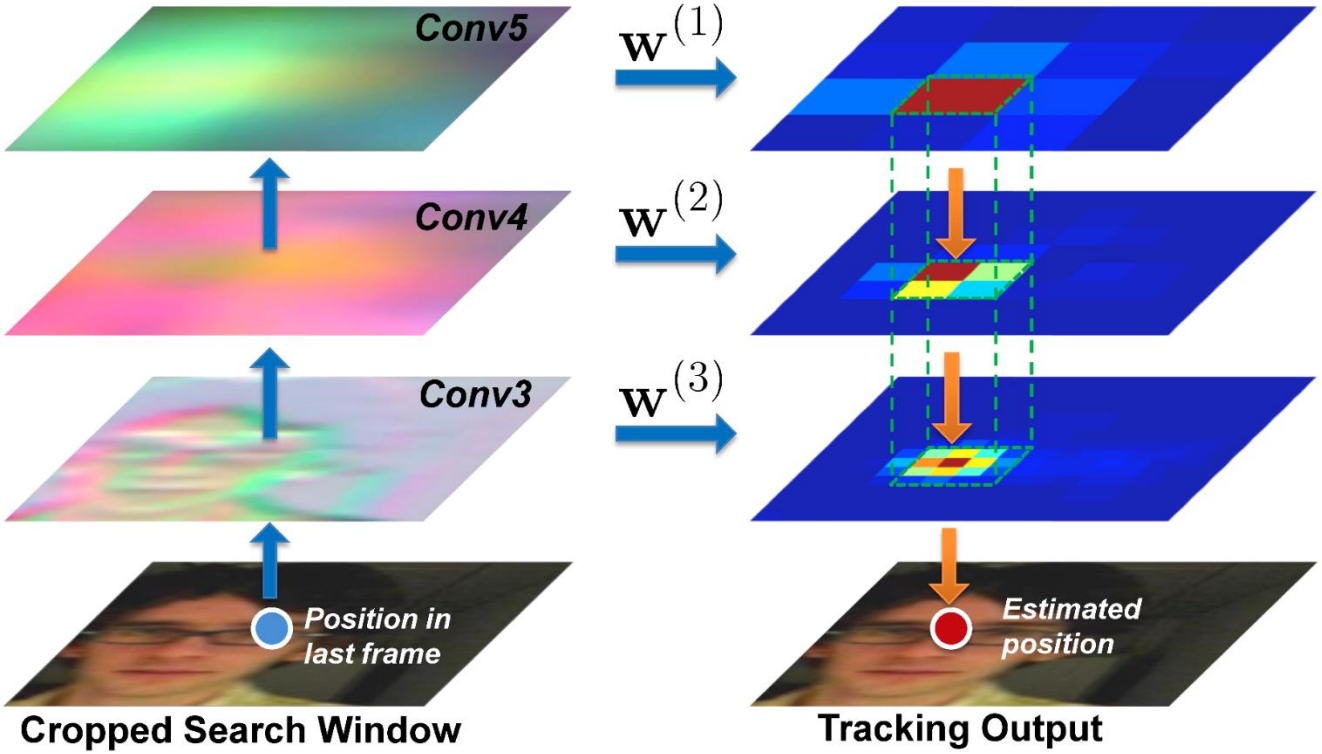


- Layer *conv5* robust to appearance change: insensitive to the sharp step edge
- Layer *conv3* is useful for precise localization: sensitive to the edge position

Feature Visualization using VGG-Net-19



Flowchart of Our Approach



Issues of Existing CNN Trackers

- Only use the last (fully-connected) layer of the CNN network for classification
  - Too coarse to localize target precisely
- Sample target states with binary labels (positive and negative)
  - Ambiguity in labeling the spatially over-correlated samples
    - MDNet (CVPR'16): negative mining
    - Struck (ICCV'11): structure output

Alleviating Sampling Ambiguity

- Adaptive correlation filters regress the deep features with soft labels decaying from 1 to 0

  - Computational efficiency using FFT
    - Convolutional theorem: convolutional filter? correlation filter?
  - Best exploit the contextual cues
    - K. Zhang et al, Fast Visual Tracking via Dense Spatio-Temporal Context Learning, in ECCV'14

Correlation Filters

- Correlation filters learning in the spatial domain:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{m,n} \|\mathbf{w} \cdot \mathbf{x}_{m,n} - y_{m,n}\|^2 + \lambda \|\mathbf{w}\|^2$$

Vertical circular shifts of input x with corresponding soft labels generated by a Gaussian function. The first five figures credit to the KCF tracker by Henrisque et al.

- Use FFT to learn correlation filter in the frequency domain as

$$\mathbf{W} = \frac{\mathbf{Y} \odot \bar{\mathbf{X}}}{\mathbf{X}^i \odot \bar{\mathbf{X}}^i + \lambda}$$

Implementation Details: Feature Interpolation

- Problem: deeper layers with lower spatial resolution due to the pooling
  - *pool5-4* in VGG-Net is of spatial size 7 x 7, which is 1/32 of the input image 224 x 224
- Solution: resize each CNN layers with bilinear interpolation
  - Affirm that deconvolution is usually helpful for finer position inference
  - Different conclusion without feature interpolation
    - M. Danelljan et al. Convolutional Features for Correlation Filter Based Visual Tracking. In ICCV 2015 workshop

Coarse-to-Fine Inference

- For the *l-th* CNN layer with channel *D*, the response map is:

$$f_l = \mathcal{F}^{-1}\left(\sum_{d=1}^{D} \mathbf{W}^d \odot \bar{\mathbf{Z}}^d\right)$$

- Given the location $(\hat{m}, \hat{n}) = \arg\max_{m,n} f_l(m,n)$, locate the target in the (*l-1*)-th layer:

$$\begin{aligned} \arg\max_{m,n} \quad & f_{l-1}(m,n) + \gamma f_l(m,n) \\ \text{s.t.} \quad & |m - \hat{m}| + |n - \hat{n}| \leq r \end{aligned}$$

Model Update

- Use a moving average scheme to update the numerator and denominator of $\mathbf{w}^d$ separately as:

$$\mathbf{A}_t^d = (1-\eta)\mathbf{A}_{t-1}^d + \eta \mathbf{Y} \odot \bar{\mathbf{X}}_t^d; \quad (1)$$

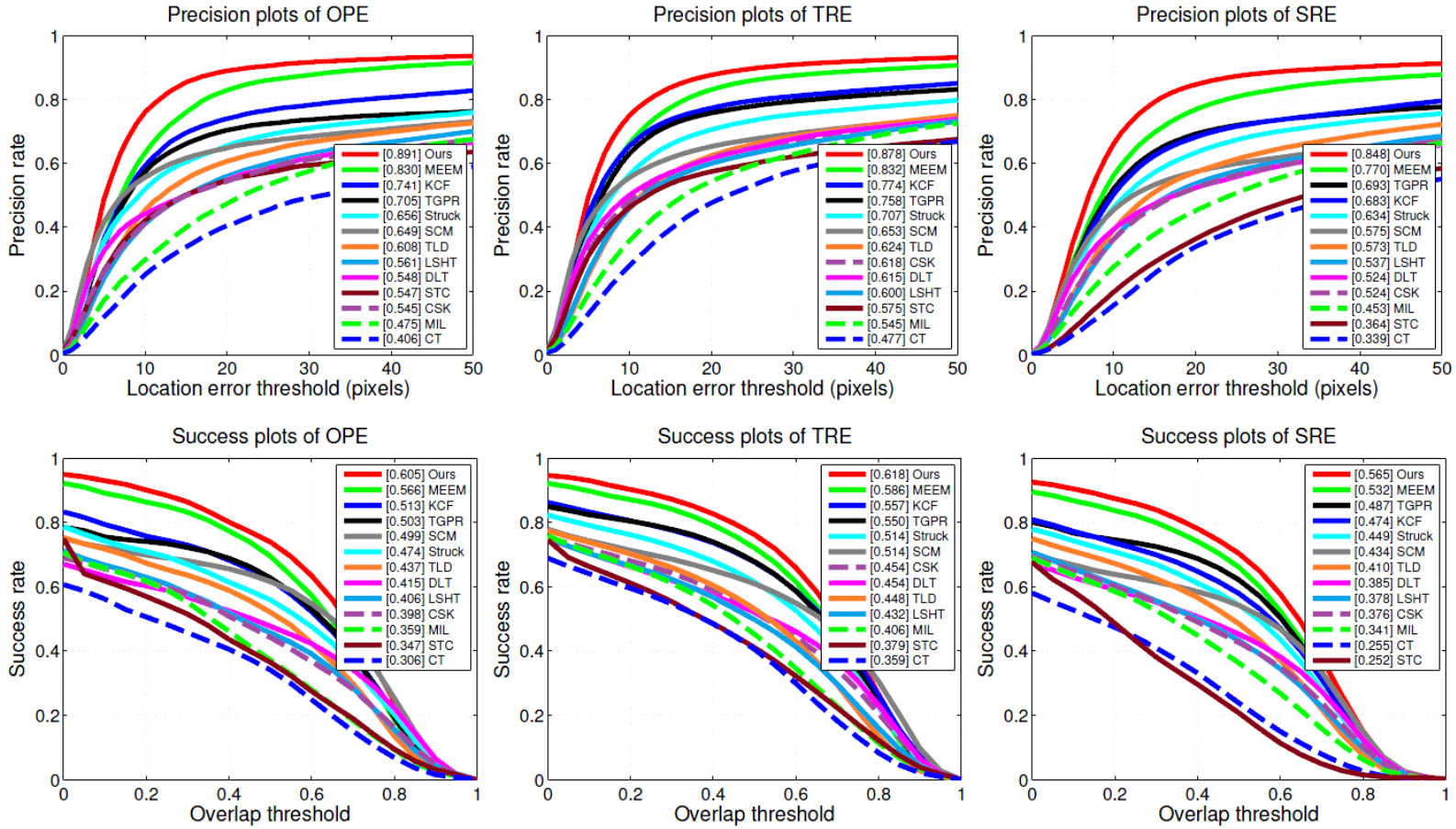
$$\mathbf{B}_t^d = (1-\eta)\mathbf{B}_{t-1}^d + \eta \sum_{i=1}^{D} \mathbf{X}_t^i \odot \bar{\mathbf{X}}_t^i; \quad (2)$$

$$\mathbf{W}_t^d = \frac{\mathbf{A}_t^d}{\mathbf{B}_t^d + \lambda}, \quad (3)$$

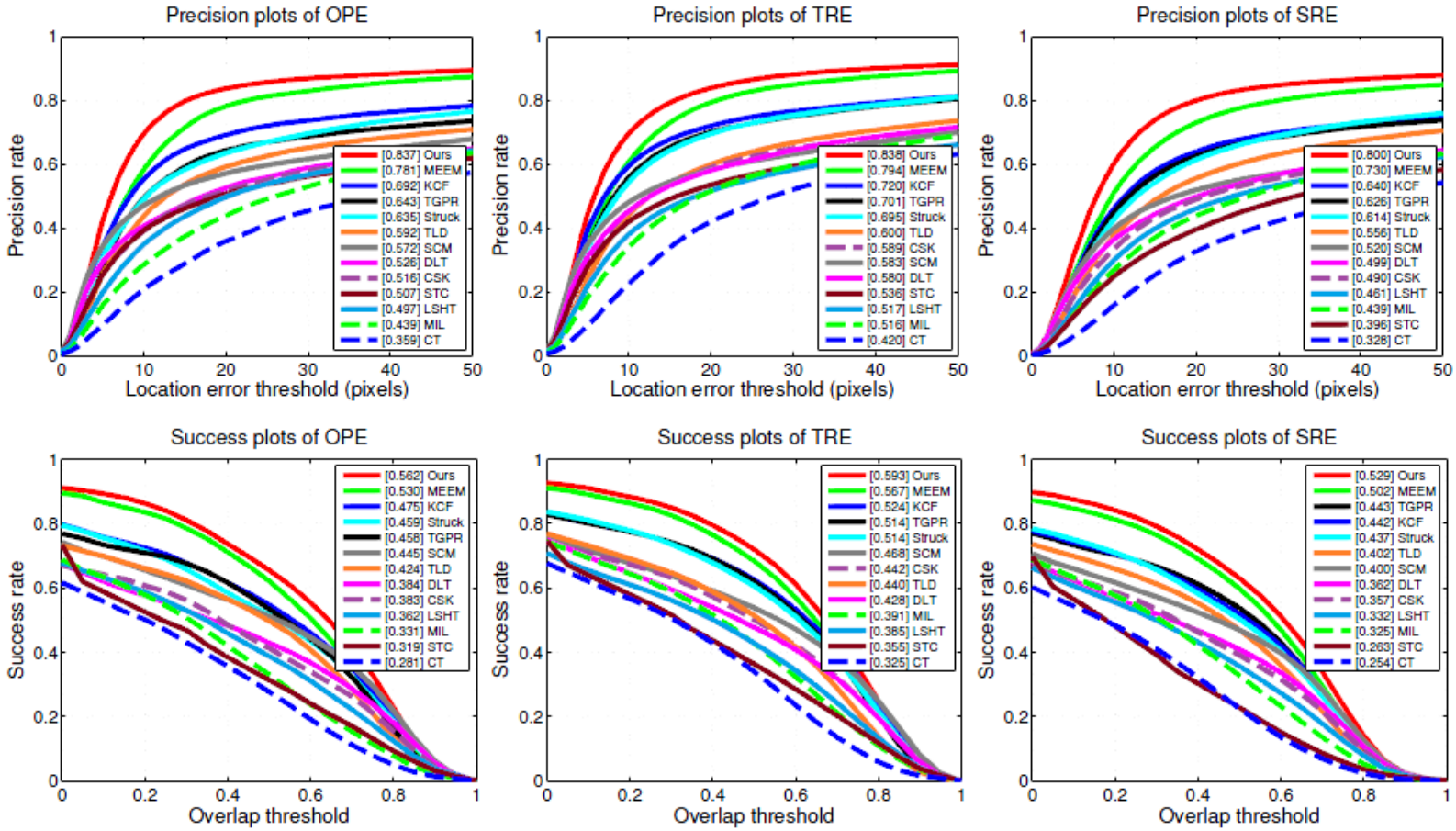
Experimental Setting

- Datasets: OTB-50, and OTB-100
  - Yi Wu et al, Online Object Tracking: A Benchmark, in CVPR, 2013
  - Yi Wu et al, Object Tracking Benchmark, TPAMI, 2015
- Metrics:
  - Distance precision rate
  - Overlap success (intersection of union) rate
- Validation schemes:
  - OPE: one-pass evaluation
  - TRE: temporal robustness evaluation
  - SRE: spatial robustness evaluation
- Fix parameters for all sequences

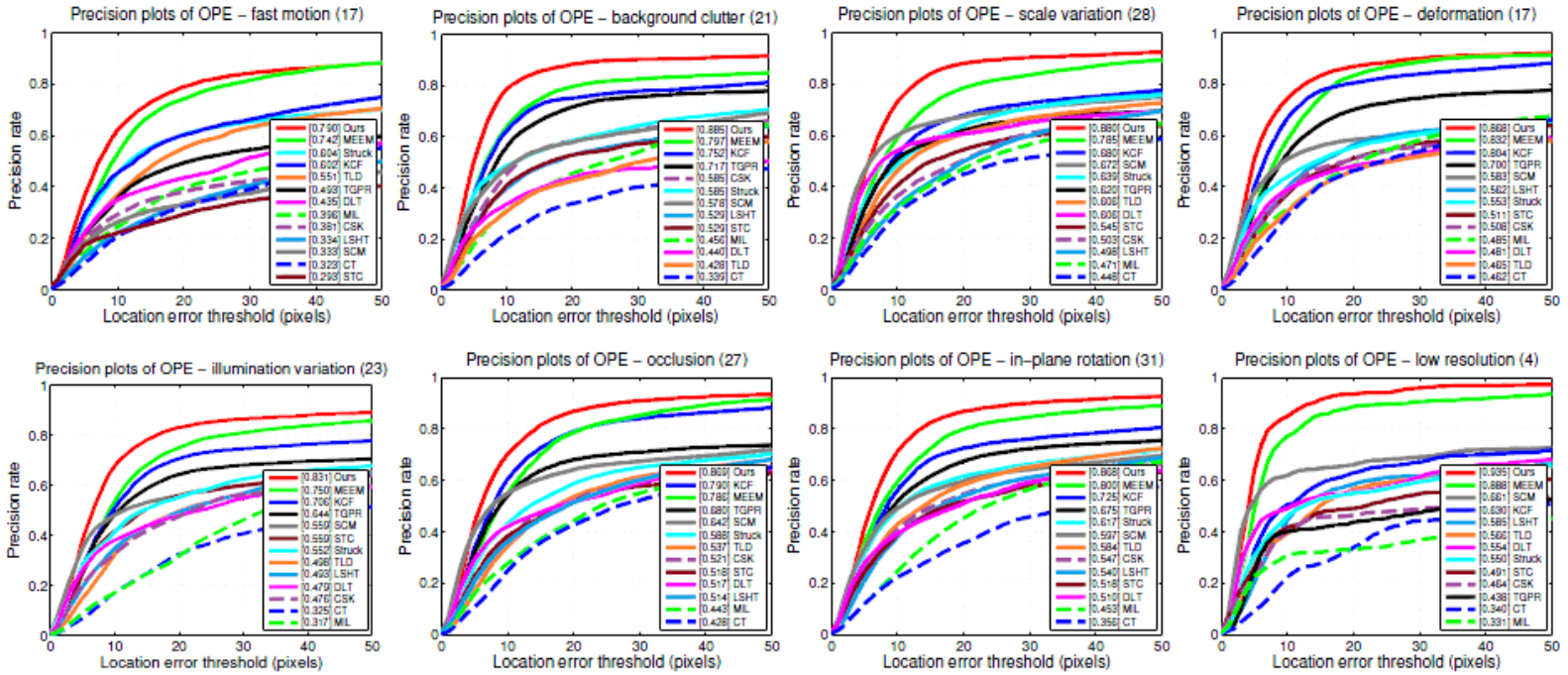
Overall Results on OTB-50



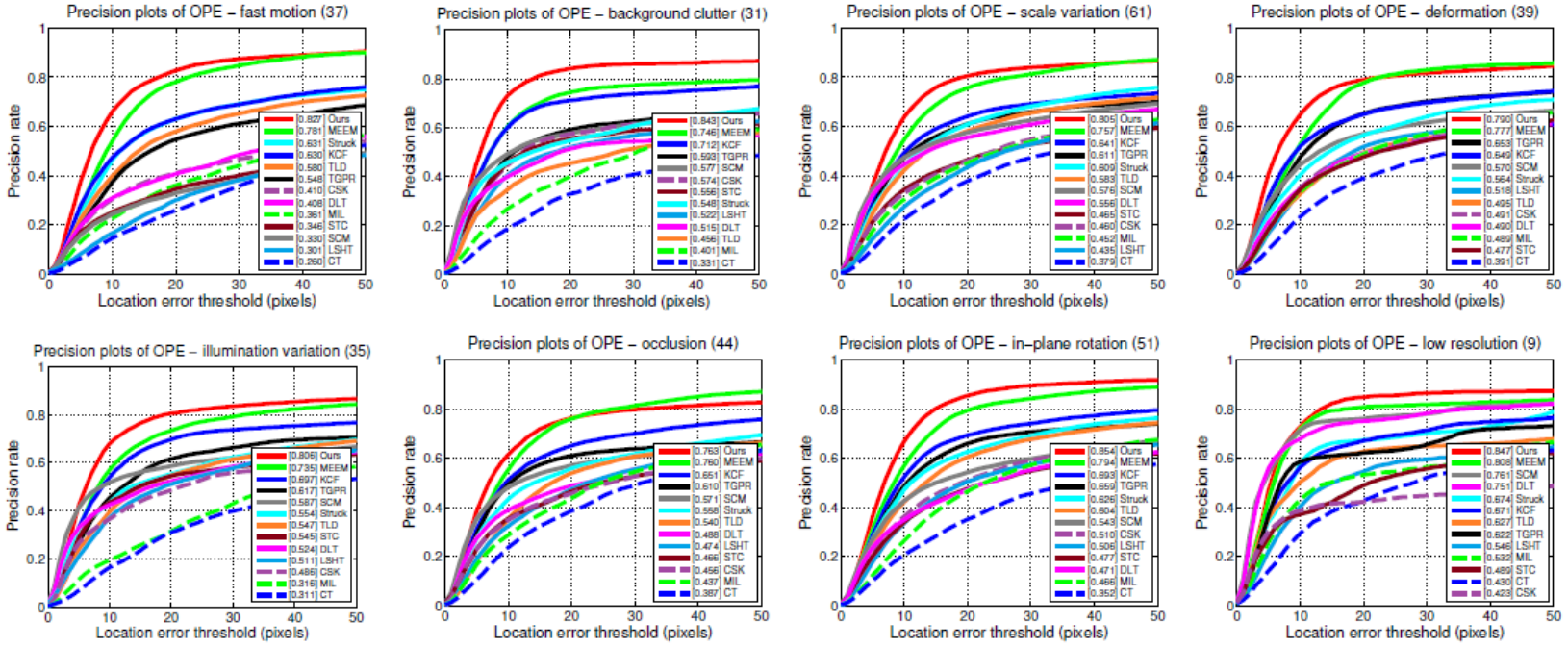
Overall Results on OTB-100



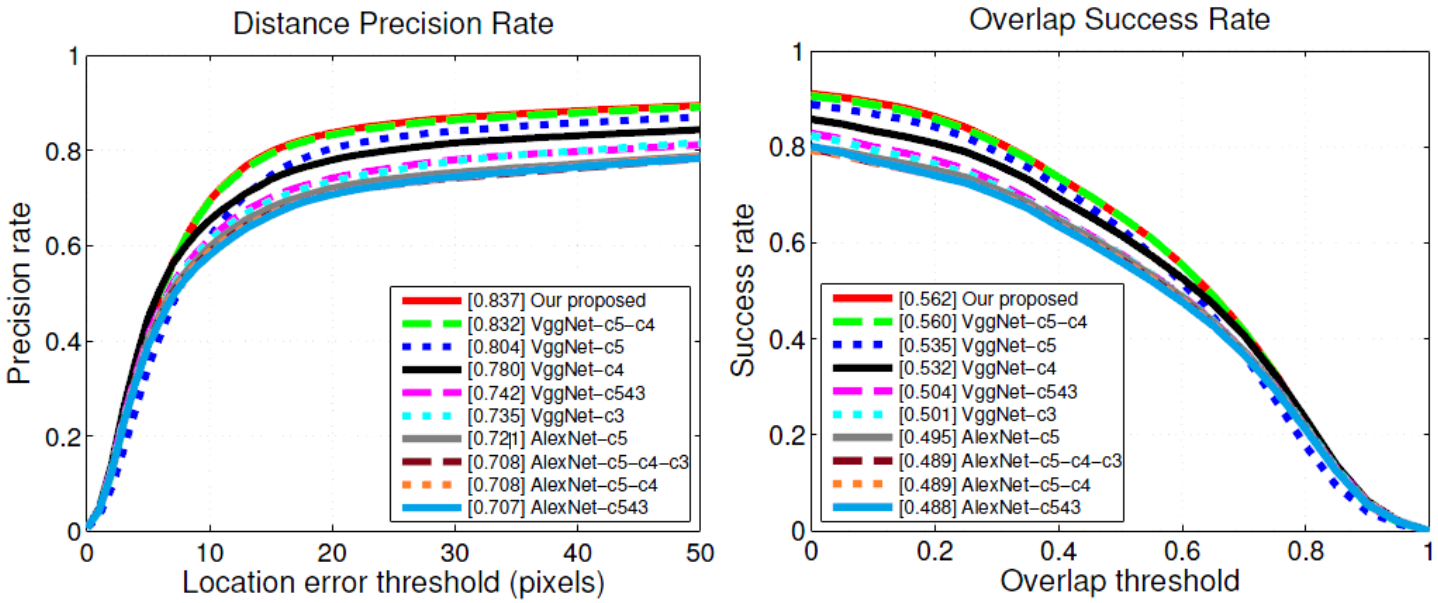
Attribute Evaluation on OTB-50



Attribute Evaluation on OTB-100

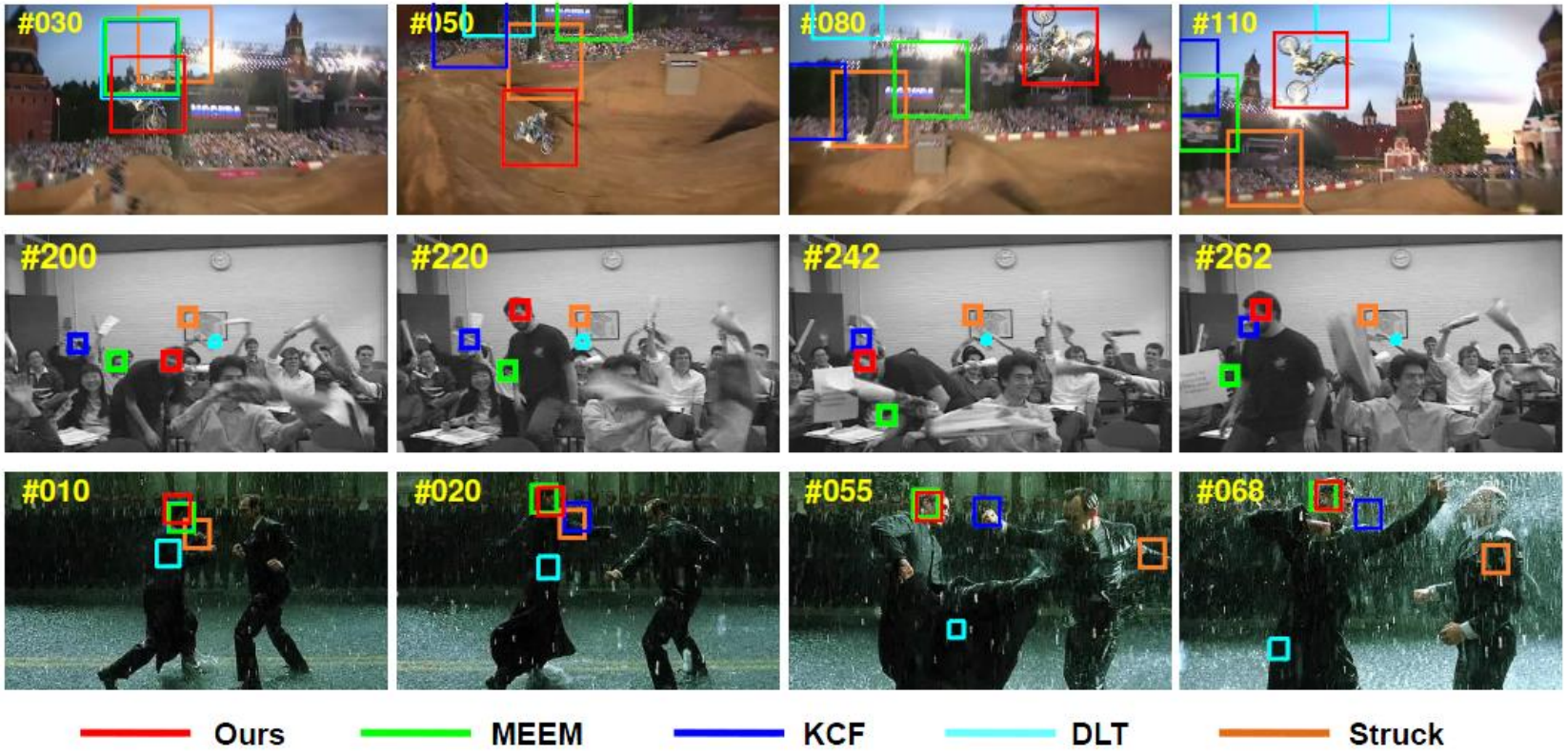


Ablation Studies

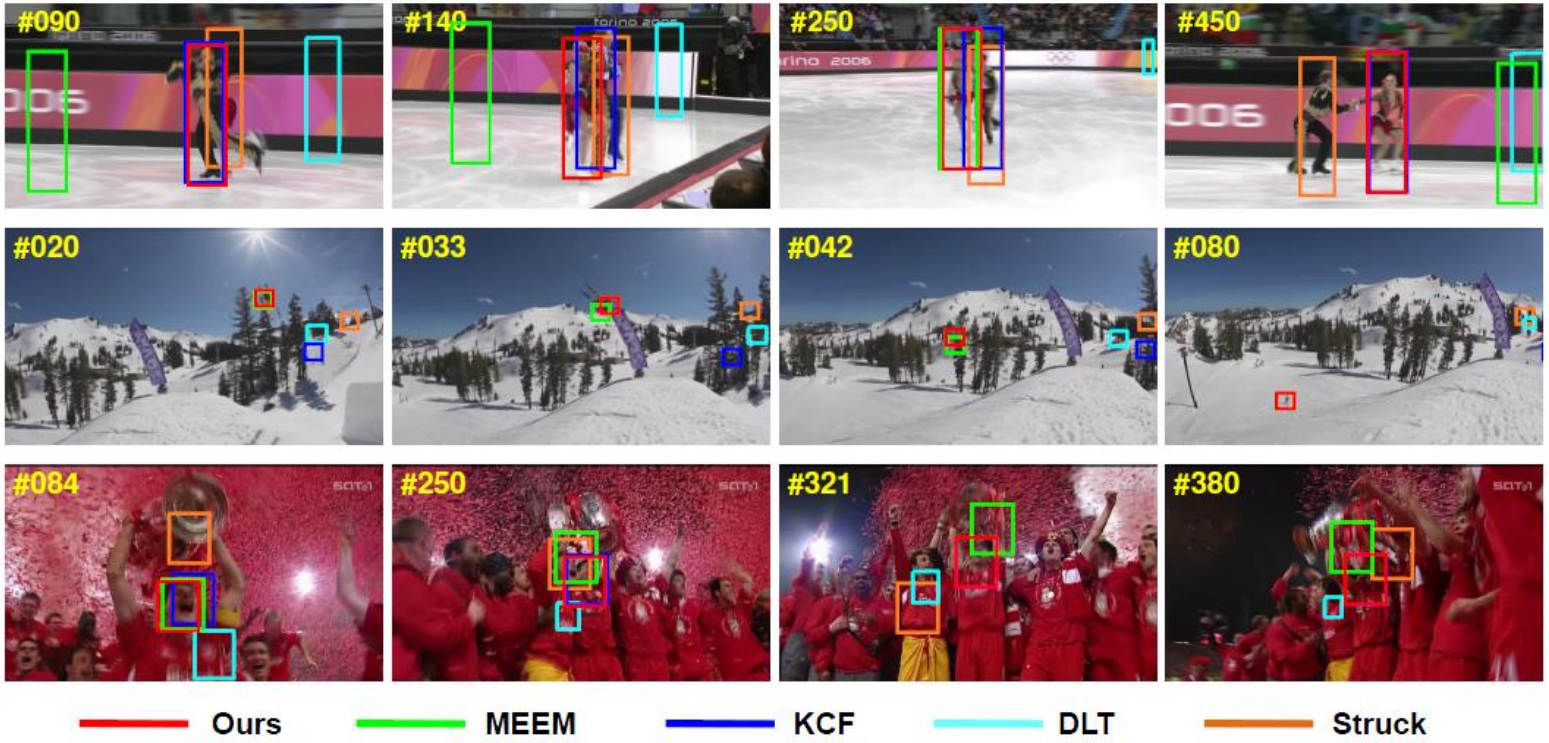


- Single layer (c5,c4 and c3), combination of the conv5-4 and conv4-4 layers (c5-c4), and concatenation of three layers (c543)

Qualitative Results I



Qualitative Results II



Failure Cases

Red boxes: our results; green: ground truth

Public Sources on This Work

- Project webpage
  - https://sites.google.com/site/chaoma99/iccv15_tracking
- Source code
  - https://github.com/jbhuang0604/CF2
- Further release the results of nine baseline trackers on OTB-100
  - https://sites.google.com/site/chaoma99/iccv15_tracking

Thanks