Attentional Correlation Filter Network for Adaptive Visual Tracking

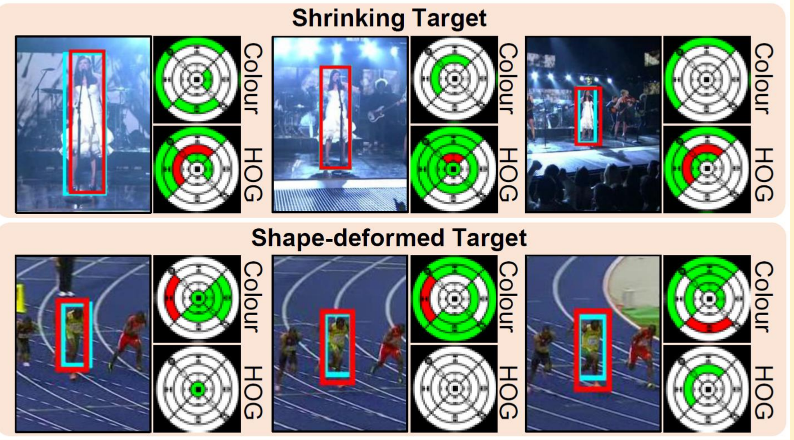
Jongwon Choi [1], Hyung Jin Chang [2], Sangdoo Yun [1], Tobias Fischer [2], Yiannis Demiris [2], Jin Young Choi [1]
jwchoi.pil@gmail.com, {yunsd101, jychoi}@snu.ac.kr, {hj.chang, t.fischer, y.demiris}@imperial.ac.uk
[1]Dept. of EC. Eng., ASRI, Seoul National Univ., South Korea. [2]Dept. of EE. Eng., Imperial College London, UK.

Codes & Results are available.
homepage: https://sites.google.com/site/jwchoivision

Target Problems

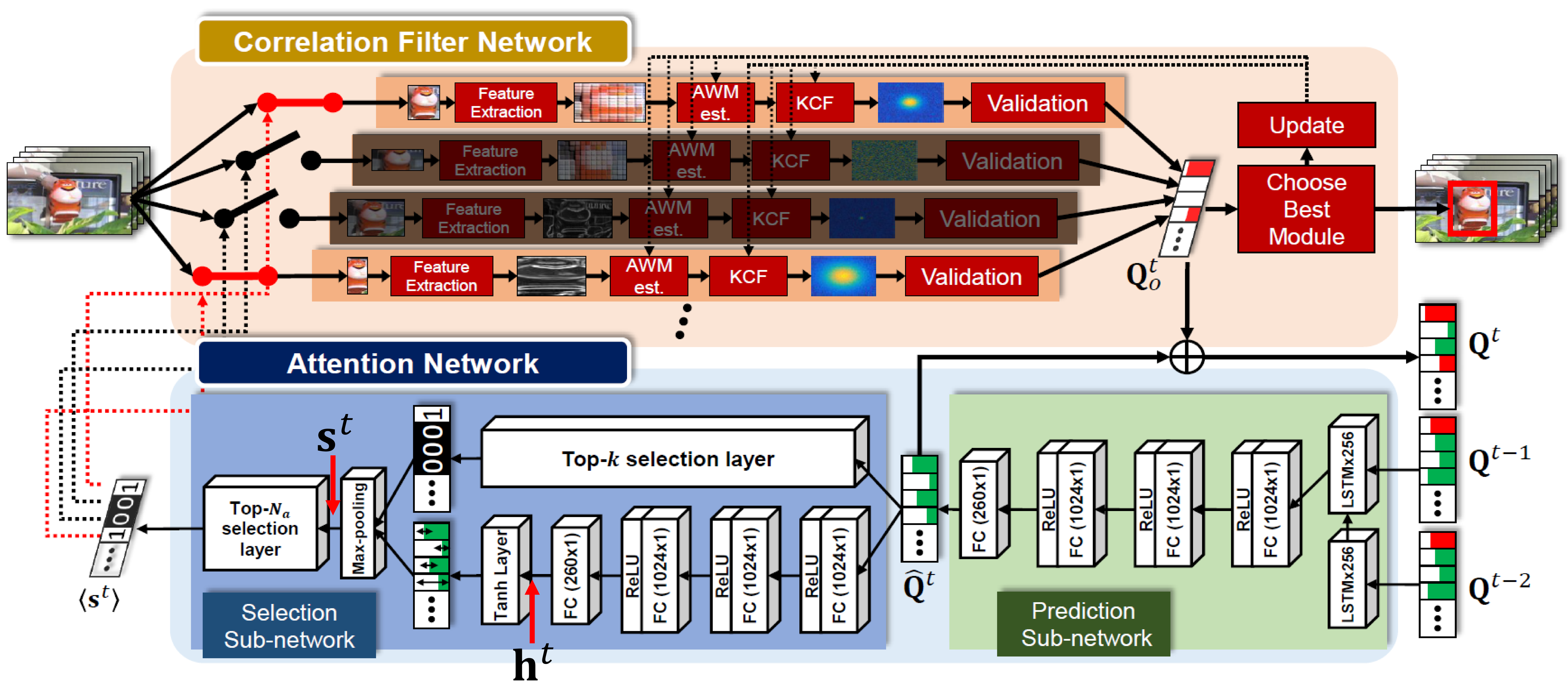
- By using many properties, tracking performance can be improved
- But, needs much time to consider various properties of target

Approach & Contribution



- **Attentional Correlation Filter Network**
  - Attention Network
    - >> Predict the module-wise performance
    - >> Select the attentional modules
  - Correlation Filter Network
    - >> A lot of tracking modules with different properties
    - >> Novel properties (flexible aspect ratio, delay etc.)

Overall Framework



Tracking Step

Score Prediction → Select Active Modules → Run Active Modules → Validate Modules → Track by Best Module → Update All Modules

Score Prediction

From prev. score vectors,

$$\{\mathbf{Q}^{t-1}, \mathbf{Q}^{t-2}, \dots\}$$

Prediction sub-network

$$\widehat{\mathbf{Q}}^t \in \mathbb{R}^{260}$$

predict curr. score vector

Select Active Modules

Selection sub-network

- Two Conditions
  - High predicted validation scores
  - High prediction error on score prediction

Run Active Modules / Validate Modules

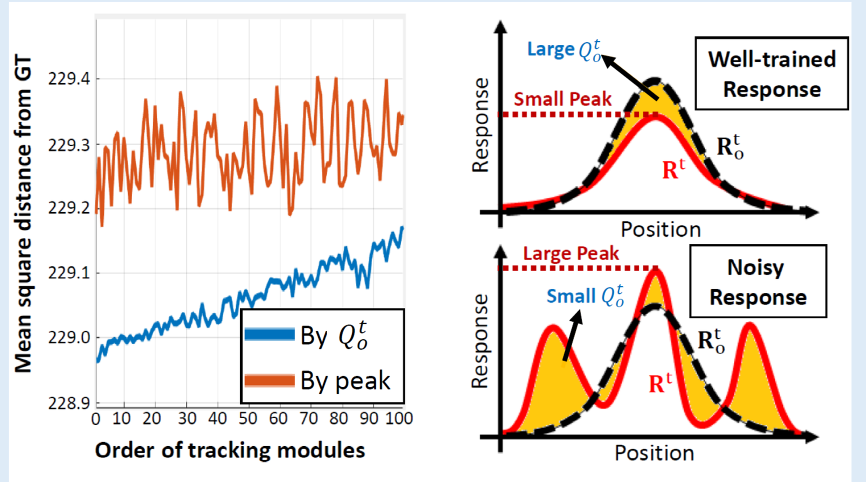
- Validation Score for Active Modules
  - Use Euclidian distance to ideal response

$$Q_o^t = \exp(-\|\mathbf{R}^t - \mathbf{R}_o^t\|_2^2)$$

$$\mathbf{R}_o^t = \mathcal{G}(\mathbf{p}'^t, \sigma_G^2)_{W \times H}$$

- Predicted Score for Inactive Modules

$$\mathbf{Q}^t = (1 - \langle \mathbf{s}^t \rangle) * \widehat{\mathbf{Q}}^t + \langle \mathbf{s}^t \rangle * \mathbf{Q}_o^t$$



(a) Reliability of validation scores (b) Validation score estimation

Update All Modules

- Only a part of modules
  - Different Feature
  - Different Kernel
- Scale change
  - Share non-scalable CF
- Delayed update
  - Reuse previous CFs

Correlation Filter Network

- **260 Tracking Modules**
  - Each tracking module is AtCF [1]
  - 2 Features (Color intensity, HOG)
  - 2 Kernel types (Gaussian, Polynomial)
  - 13 Flexible scale changes (-2x, -x, +x, +2x, -2y, -y, +y, +2y, +xy, +2xy, 0)
  - 5 Delayed updates (0, -1, -2, -3, -4 frames)

Pre-training of Attention Network

- **Loss Function**

$$E = \sum_{i=1}^{N} \left\{ \|\mathbf{Q}(i) - \mathbf{Q}_{GT}(i)\|_2^2 + \lambda \|\mathbf{s}(i)\|_0 \right\}$$
$$\mathbf{Q}(i) = (\mathbf{1} - \langle \mathbf{s}(i) \rangle) * \mathbf{Q}(i) + \langle \mathbf{s}(i) \rangle * \mathbf{Q}_{GT}(i)$$

Relaxation

$$E = \sum_{i=1}^{N} \left\{ \left\| (\mathbf{1} - \mathbf{s}(i)) * (\widehat{\mathbf{Q}}(i) - \mathbf{Q}_{GT}(i)) \right\|_2^2 + \lambda \|\mathbf{s}(i)\|_0 \right\}$$
$$\mathbf{Q}(i) = (\mathbf{1} - \mathbf{s}(i)) * \widehat{\mathbf{Q}}(i) + \mathbf{s}(i) * \mathbf{Q}_{GT}(i)$$

- Prediction sub-network

$$E = \sum_{i=1}^{N} \left\{ \left\| \widehat{\mathbf{Q}}(i) - \mathbf{Q}_{GT}(i) \right\|_2^2 \right\}$$

- Selection sub-network

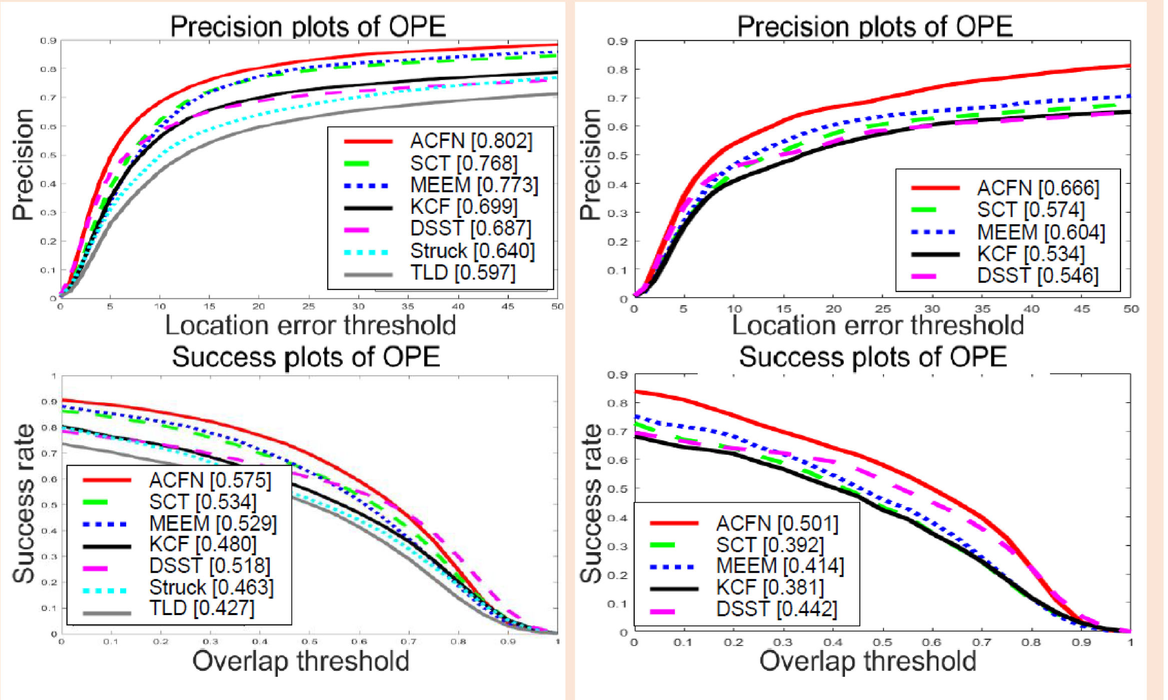
$$E = \sum_{i=1}^{N} \left\{ \left\| (\mathbf{1} - \mathbf{s}(i)) * (\widehat{\mathbf{Q}}(i) - \mathbf{Q}_{GT}(i)) \right\|_2^2 + \lambda \ln(1 + \|\mathbf{h}(i)\|_1) \right\}$$

Experiment

- **Implementation**
  - Tensorflow (CF-Net) + MATLAB (At-Net) (By socket communication)
  - i7-6900K CPU, 32GB RAM, NVIDIA GTX1070 CPU
- **Quantitative Results**

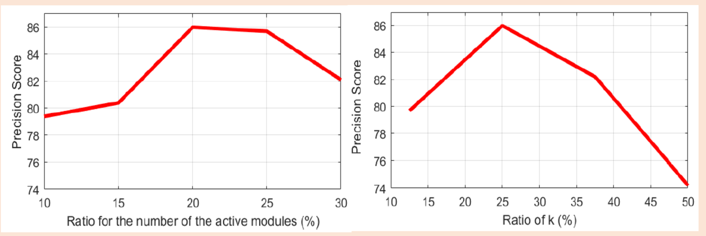
Table 1. Quantitative results on the CVPR2013 dataset [38]

	Algorithm	Pre. score	Mean FPS	Scale
Proposed	ACFN	86.0%	15.0	O
	CFN+predNet	82.3%	14.4	O
	CFN	81.3%	6.9	O
	CFN+simpleSel.	79.4%	15.7	O
	CFN-	78.4%	15.5	O
Real-time	SCT [3]	84.5%	40.0	X
	MEEM [42]	81.4%	19.5	X
	KCF [16]	74.2%	**223.8**	X
	DSST [5]	74.0%	25.4	O
	Struck [15]	65.6%	10.0	O
	TLD [19]	60.8%	21.7	O
Non Real-time	C-COT [8]	**89.9%**	<1.0	O
	MDNet-N [29]	87.7%	<1.0	O
	MUSTer [18]	86.5%	3.9	O
	FCNT [35]	85.6%	3.0	O
	D-SRDCF [6]	84.9%	<1.0	O
	SRDCF [7]	83.8%	5	O
	STCT [36]	78.0%	2.5	O

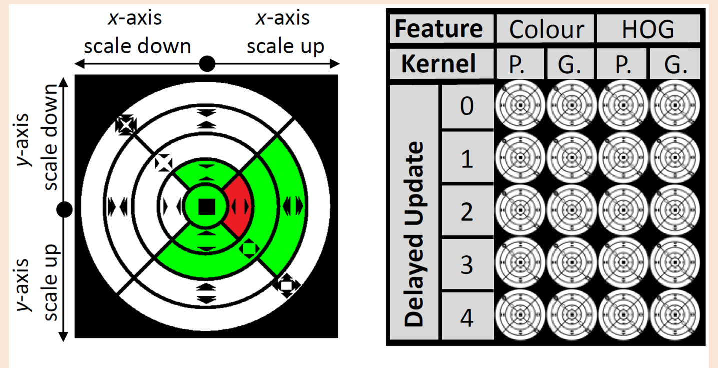


(b) Evaluation plots on TPAMI2015 dataset (c) Evaluation plots on VOT2014 dataset

- **Analysis**
  - Parameter Analysis

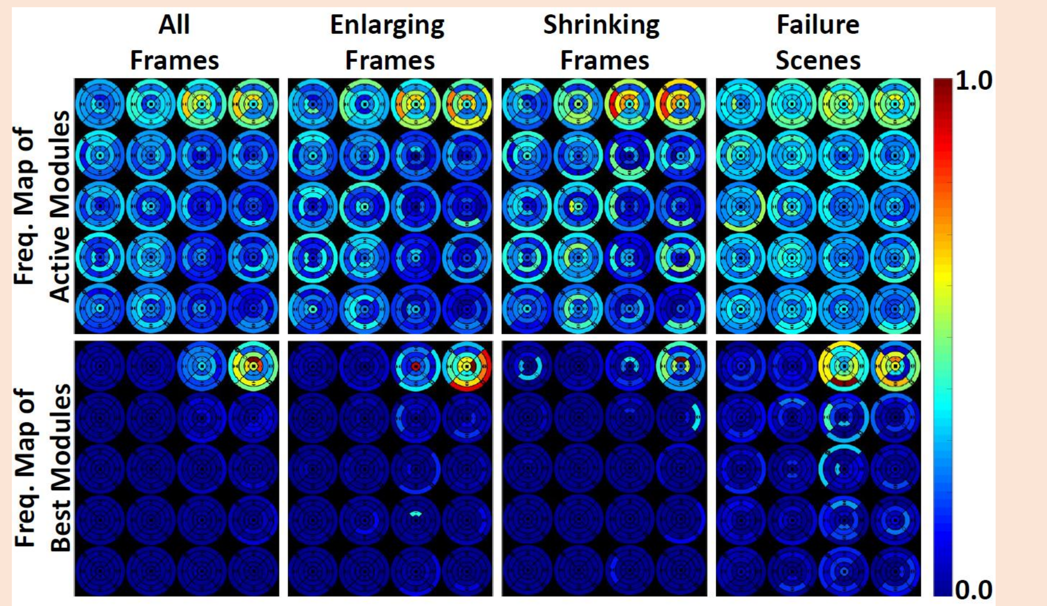


  - Attention Map Definition

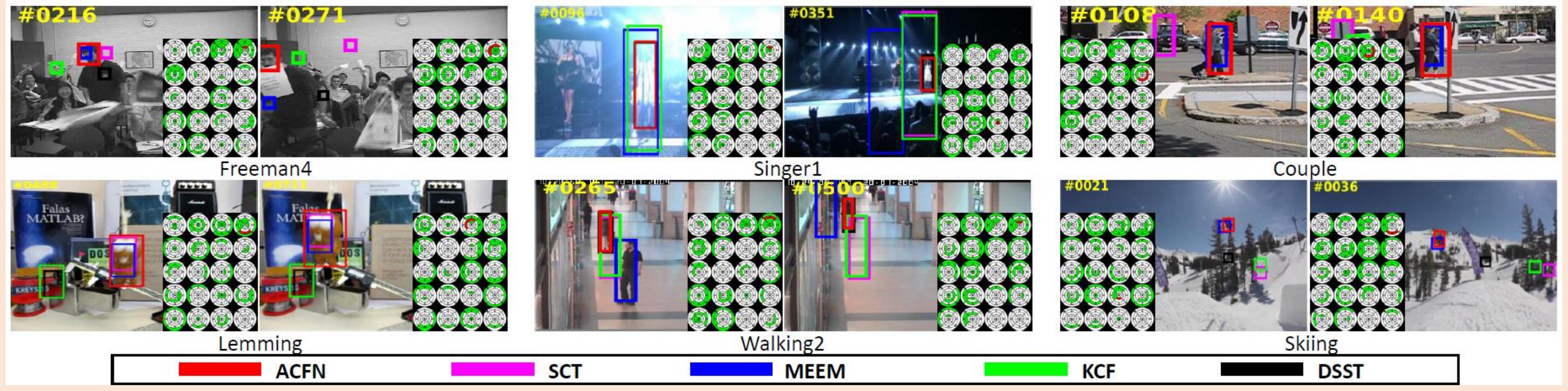


(a) Attention map (b) Global attention map

  - Frequency map for various cases



- **Qualitative Results**



Reference

[1] Choi et al., "Visual tracking using attention-modulated disintegration and integration", CVPR2016

This work was partly supported by the ICT R&D program of MSIP/IITP (No.B0101-15-0552, Development of Predictive Visual Intelligence Technology), the SNU-Samsung Smart Campus Research Centre, Brain Korea 21 Plus Project, and EU FP7 project WYSIWYD under Grant 612139.