# Hierarchical Attentive Recurrent Tracking

Adam R. Kosiorek, Alexy Bewley, Ingmar Posner

Applied AI Lab, Department of Engineering Science, University of Oxford, UK

## Problem Statement

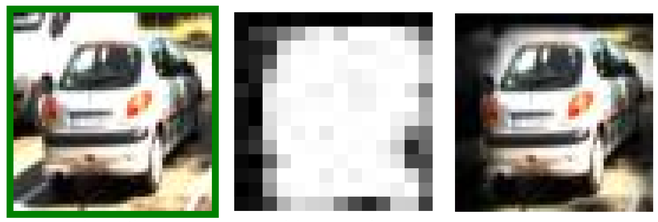
**What:** Class-agnostic single object tracking in real-world videos with camera motion

**Difficulties:** No target-specific discriminative models
Cluttered backgrounds with many distractors

**How:** Discard uninformative background features
Learn arbitrary motion models
Anticipate appearance changes

**Approach:** Recurrent Neural Network with Hierarchical Attention Mechanism

## Hierarchical Attention



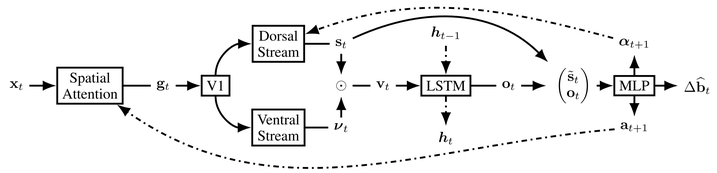

**Bio-inspired:** Two-stream processing pathway and attention mechanisms adapted from human visual cortex.

**Interpretable:** Important features selected by Spatial Attention and Object Segmentation mechanisms.

**Scalable:** Applicable to real-world data due to distractor suppression and auxiliary loss terms.

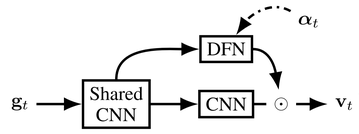
**Efficient:** Attention quickly discards irrelevant features $> 120$ fps on a laptop!

## Two-Stream Attentive Model




| | | | |
|---|---|---|---|
| $\mathbf{x}_t$ | input image | $\boldsymbol{h}_t$ | hidden state |
| $\mathbf{g}_t$ | attention glimpse | $\mathbf{o}_t$ | LSTM output |
| $\boldsymbol{\nu}_t$ | appearance-based features | $\boldsymbol{\alpha}_{t+1}$ | appearance |
| $\mathbf{s}_t$ | object segmentation | $\Delta\widehat{\mathbf{b}}_t$ | bounding-box update |
| $\mathbf{v}_t$ | masked features | $\mathbf{a}_{t+1}$ | spatial attention |

Appearance attention architecture:




**Dorsal Stream:** Dynamic Filter Network [4] conditioned on appearance finds target-specific features

**V1:** Shared CNN **Ventral Stream:** CNN extracts visual features

## Loss

Directly optimise Intersection-over-Union (IoU) and guide attention mechanisms.

$$\mathcal{L}_{\text{HART}}(\cdot) = \lambda_{\text{t}}\mathcal{L}_{\text{t}}(\cdot) + \lambda_{\text{s}}\mathcal{L}_{\text{s}}(\cdot) + \lambda_{\text{a}}\mathcal{L}_{\text{a}}(\cdot) + \beta R(\cdot)$$

**Tracking:** Negative log of Intersection-over-Union.

$$\mathcal{L}_{\text{t}}(\mathcal{D}, \theta) = \mathbb{E}_{p(\widehat{\mathbf{b}}_{1:T}|\mathbf{x}_{1:T}, \mathbf{b}_1)}\left[-\log \text{IoU}(\widehat{\mathbf{b}}_t, \mathbf{b}_t)\right]$$

**Spatial Attention:** It follows the object, but shouldn't be too big.

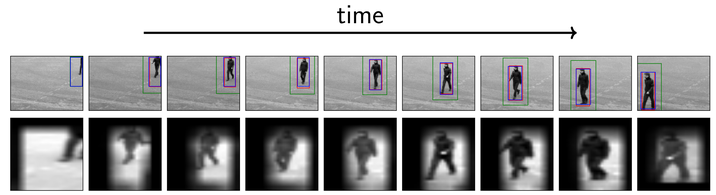
$$\mathcal{L}_{\text{s}}(\mathcal{D}, \theta) = \mathbb{E}_{p(\mathbf{a}_{1:T}|\mathbf{x}_{1:T}, \mathbf{b}_1)}\left[-\log\left(\frac{\mathbf{a}_t \cap \mathbf{b}_t}{\text{area}(\mathbf{b}_t)}\right) - \log(1 - \text{IoU}(\mathbf{a}_t, \mathbf{x}_t))\right]$$

**Appearance Attention:** Cross-entropy with dynamically created target mask $\tau(\mathbf{a}_t, \mathbf{b}_t)$: $\mathcal{L}_{\text{a}}(\mathcal{D}, \theta) = \mathbb{E}_{p(\mathbf{a}_{1:T}, \mathbf{s}_{1:T}|\mathbf{x}_{1:T}, \mathbf{b}_1)}[H(\tau(\mathbf{a}_t, \mathbf{b}_t), \mathbf{s}_t)]$.

With Appearance Attention Loss: Successful Tracking

Without Appearance Attention Loss: ID Swap

## Pedestrian Tracking: KTH Dataset [2]

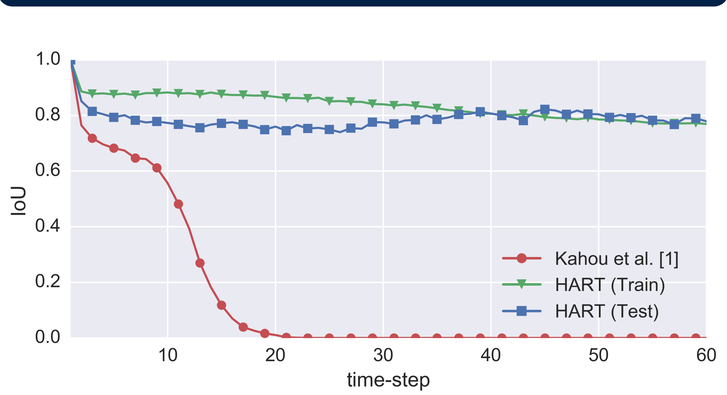



- Attention, prediction and ground-truth overlap at initialization.
- Every $16^{th}$ frame of the sequence at 25 fps.
- $2^{nd}$ row: attention glimpses multiplied with appearance attention.

Intersection over Union

| Kahou *et. al.* [1] | Ours |
|---|---|
| 0.55 | **0.77** |

## Scaling to Real-World Data: KITTI [3]




Average IoU on KITTI over 60 time-steps

| Kahou *et. al.* [1] | Spatial Att | App Att | **HART** |
|---|---|---|---|
| 0.14 | 0.60 | 0.78 | **0.81** |

Spatial Att - no appearance attention
App Att - no appearance attention loss

IoU curves on KITTI over 60 time-steps; HART (train) presents evaluation on the train set.

## References

**[1]** Samira Ebrahimi Kahou, Vincent Michalski, and Roland Memisevic. RATM: Recurrent Attentive Tracking Model. CVPR Work., 2017.

**[2]** Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In ICPR. IEEE, 2004.

**[3]** A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The KITTI dataset. Int. J. Rob. Res., 32(11):12311237, 2013.

**[4]** Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic Filter Networks. NIPS, 2016.