# Hierarchical Attentive Recurrent Tracking

Adam R. Kosiorek, Alex Bewley, Ingmar Posner

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

{adamk, bewley, ingmar}@robots.ox.ac.uk

## Problem Statement

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

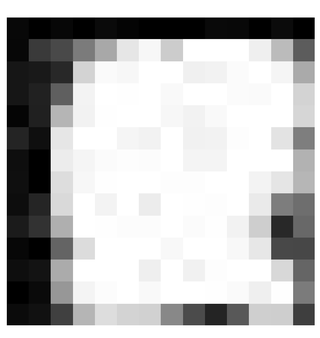
**Difficulties:** No target-specific discriminative models
Cluttered backgrounds with many distractors

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

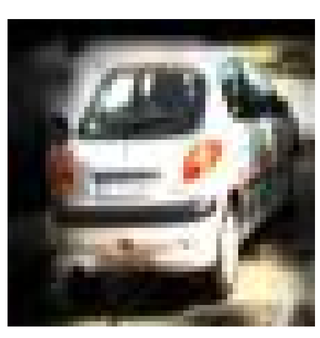
**Approach:** Recurrent Neural Network with Hierarchical Attention Mechanism

## Hierarchical Attention

attention glimpse



object mask



appearance attention

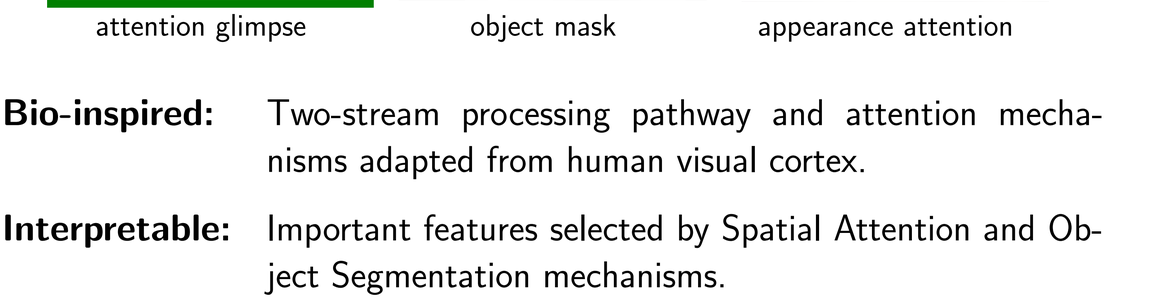
**Bio-inspired:** Two-stream processing pathway and attention mechanisms adapted from human visual cortex.

**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 |

**V1:** Shared low-level convolutional neural network for feature extraction.
**Dorsal Stream:** Attention driven by high-level appearance features in a *top-down* fashion.
**Ventral Stream:** Appearance features computed in a *bottom-up* fashion.

## Learning to Attend for Tracking

The model is optimised for the following three objectives.

**Main Tracking Objective:** Maximise the overlap of predicted box with the true object.

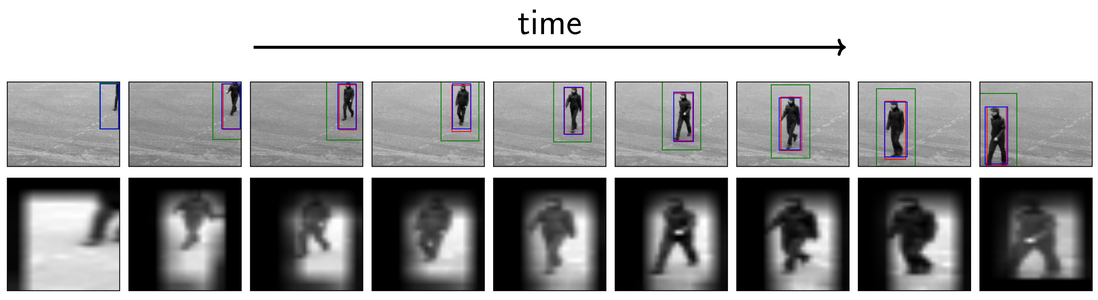
**Spatial Attention:** The glimpse should contain the object, but shouldn't be too big.

**Appearance Attention:** The dynamic appearance features should respond specifically to the target object.

Appearance attention loss (top) prevents an ID swap when a pedestrian is occluded by another one (bottom).

Left to right: glimpses and segmentations learned with and without appearance loss. Attention loss leads to distractor suppression.

## Pedestrian Tracking: KTH Dataset [2]

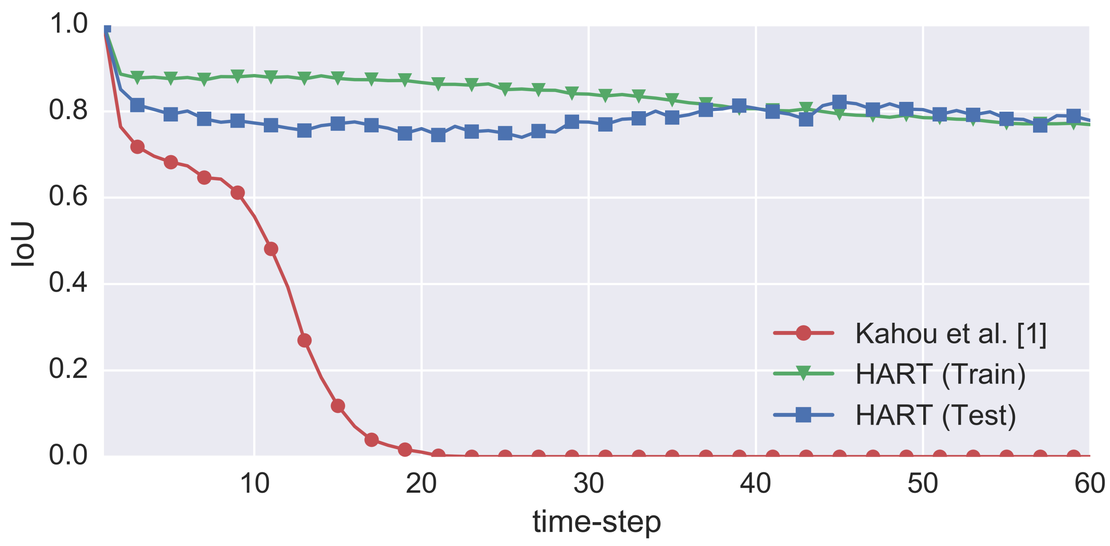



- Attention, prediction and ground-truth overlap at initialization.
- Every $16^{th}$ frame of the sequence at 25 fps.
- $2^{nd}$ row: attention glimpses multiplied with appearance attention.

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 Workshop, 2017.

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

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