

Hierarchical Attentive Recurrent Tracking

Adam R. Kosiorek, Alexy Bewley, Ingmar Posner

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

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

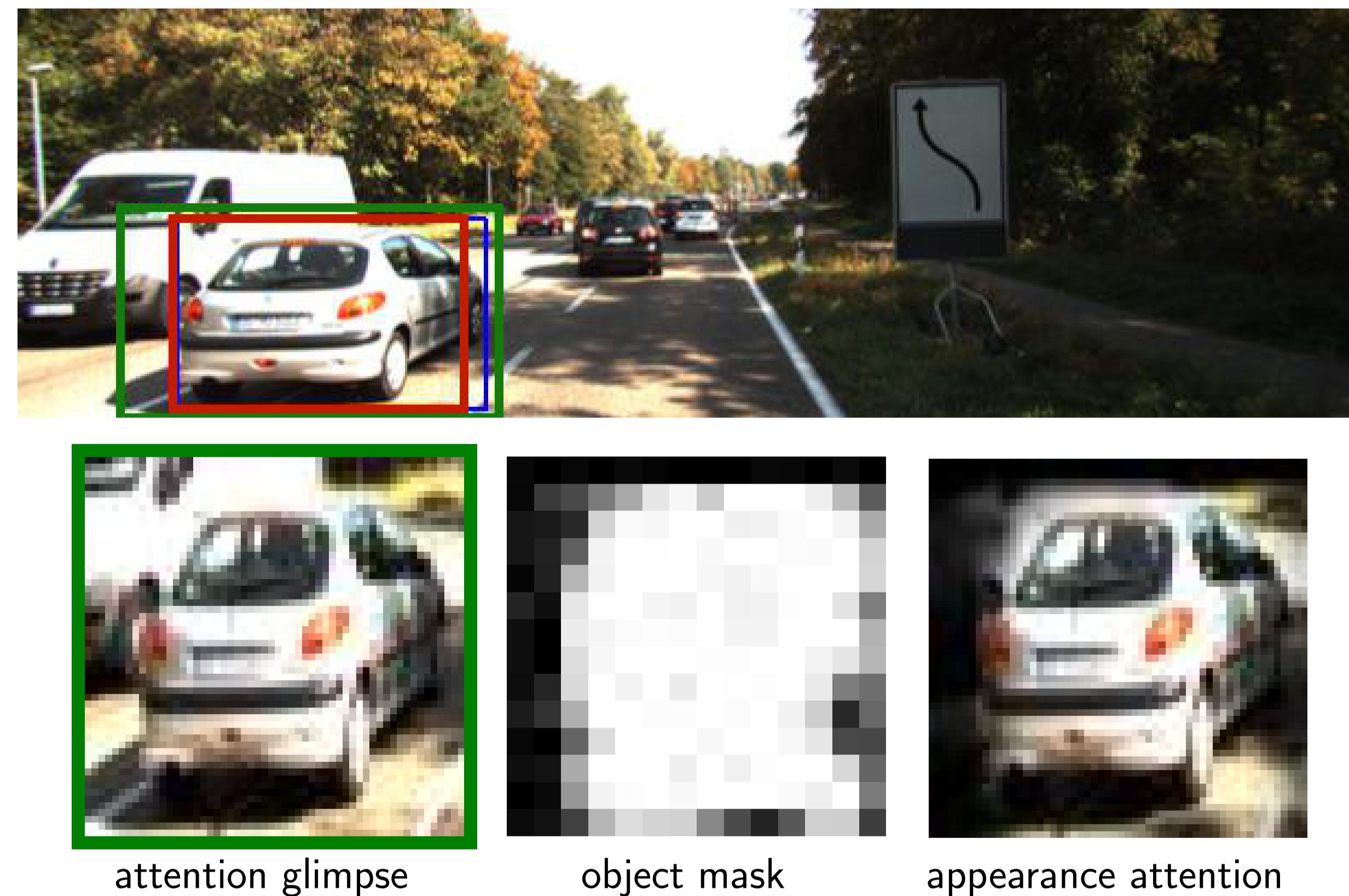


UNIVERSITY OF
OXFORD

Problem Statement

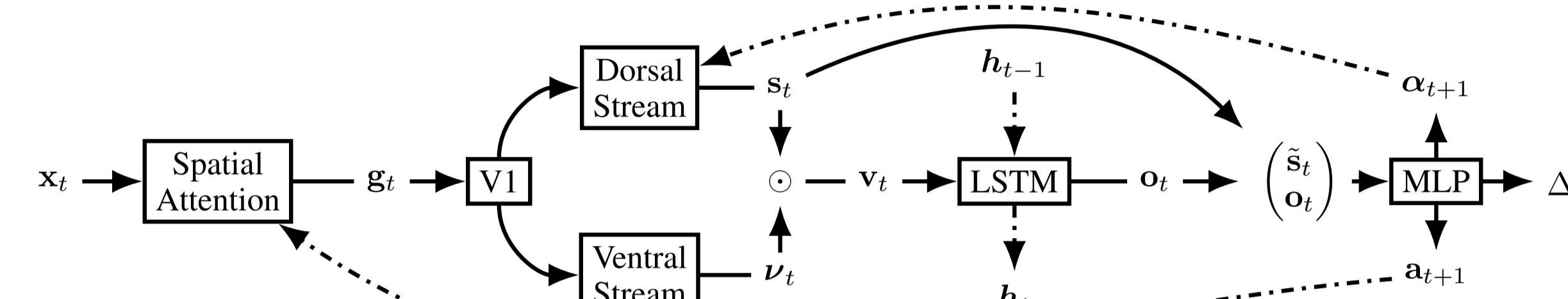
- What:** Class-agnostic single object tracking in real-world videos with camera motion
- Difficulties:** No target-specific discriminative models
Cluttered backgrounds with many distractors
- How:** Discard uninformative background features
Learn arbitrary motion models
Anticipate appearance changes
- Approach:** Recurrent Neural Network with Hierarchical Attention Mechanism

Hierarchical Attention



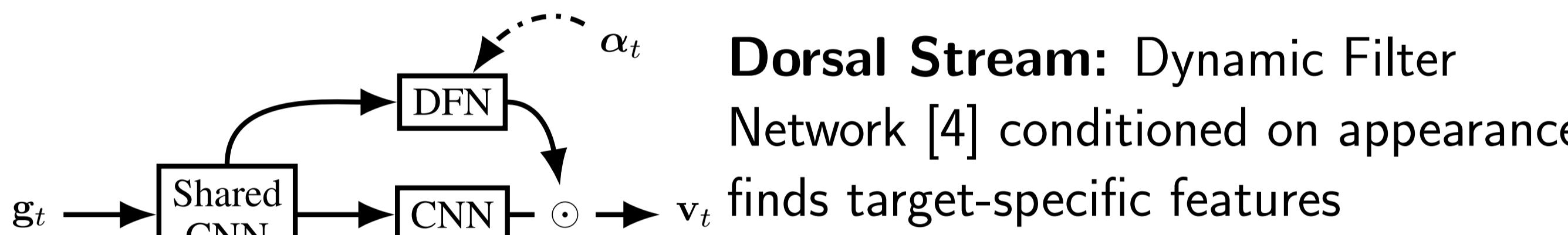
- Bio-inspired:** Two-stream processing pathway and attention mechanisms adapted from human visual cortex.
- Interpretable:** Important features selected by Spatial Attention and Object Segmentation mechanisms.
- Scalable:** Applicable to real-world data due to distractor suppression and auxiliary loss terms.
- Efficient:** Attention quickly discards irrelevant features
> 120 fps on a laptop!

Two-Stream Attentive Model



- | | | | |
|-------|---------------------------|--------------------|---------------------|
| x_t | input image | h_t | hidden state |
| g_t | attention glimpse | o_t | LSTM output |
| v_t | appearance-based features | α_{t+1} | appearance |
| s_t | object segmentation | $\Delta \hat{b}_t$ | bounding-box update |
| v_t | masked features | a_{t+1} | spatial attention |

Appearance attention architecture:



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

Loss

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

$$\mathcal{L}_{\text{HART}}(\cdot) = \lambda_t \mathcal{L}_t(\cdot) + \lambda_s \mathcal{L}_s(\cdot) + \lambda_a \mathcal{L}_a(\cdot) + \beta R(\cdot)$$

Tracking: Negative log of Intersection-over-Union.

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

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

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

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

With Appearance Attention Loss: Successful Tracking

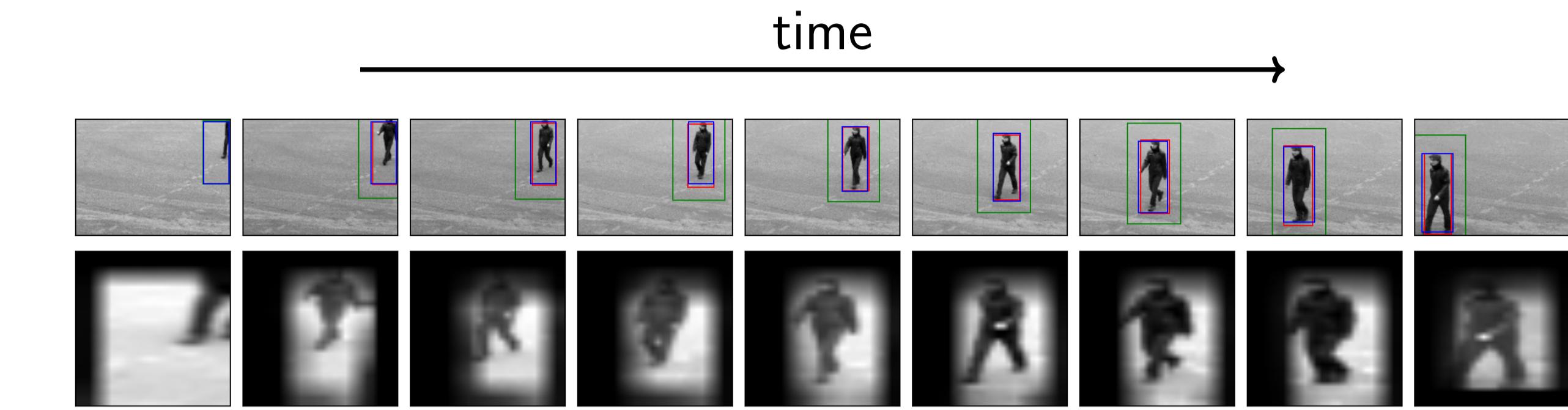


Without Appearance Attention Loss: ID Swap

Q <https://github.com/akosiorek/hart>

YouTube <https://youtu.be/Vvkjm0FRGSS>

Pedestrian Tracking: KTH Dataset [2]



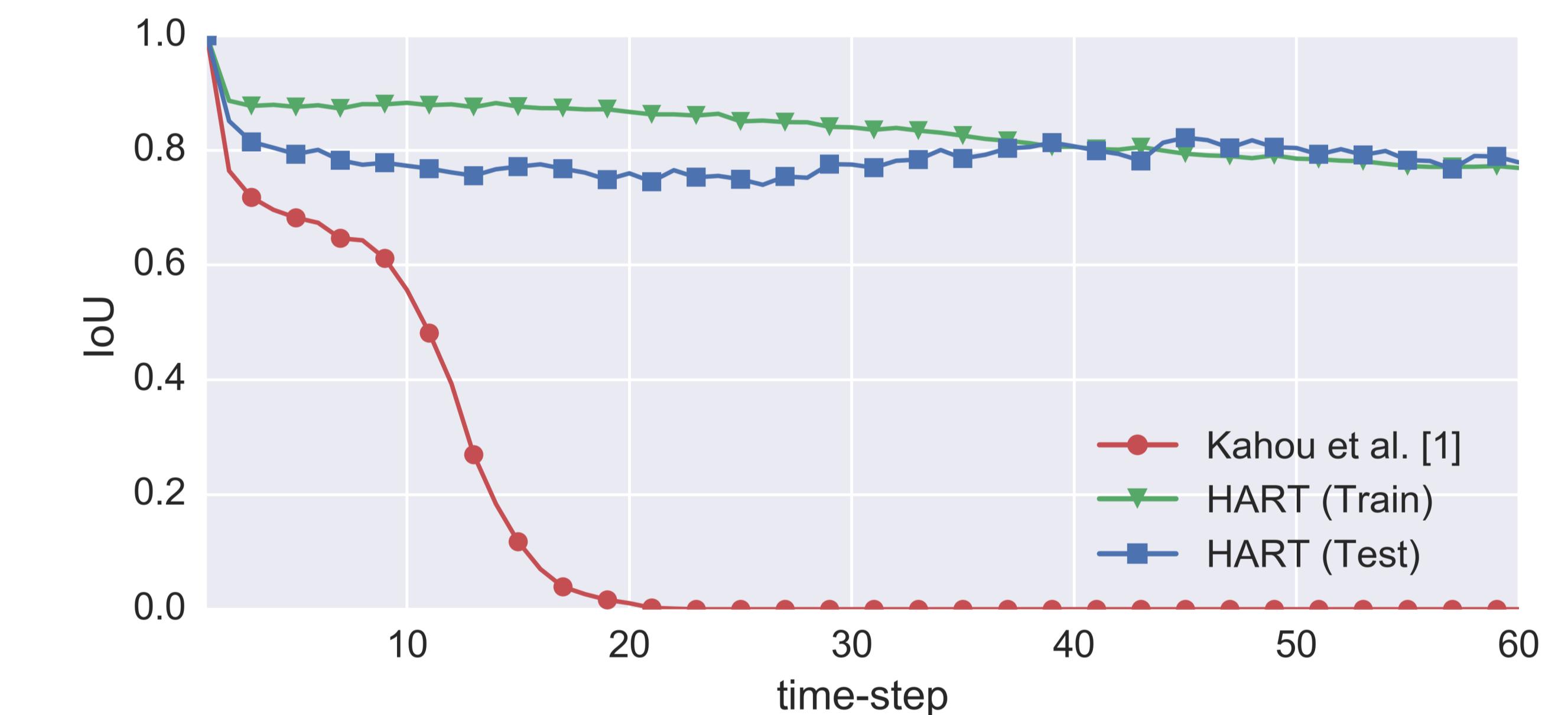
- Attention, prediction and ground-truth overlap at initialization.

- Every 16th frame of the sequence at 25 fps.

- 2nd row: attention glimpses multiplied with appearance attention.

	time
Kahou et. al. [1]	0.55
Ours	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.
- [4] Bert De Brabandere, Xu Jia, Tinne Tuytelaars, and Luc Van Gool. Dynamic Filter Networks. NIPS, 2016.