



MIDDLE EAST TECHNICAL UNIVERSITY

# MMI711

# Sequence Models

# in Multimedia

Deep Trackers  
(paper analyses)

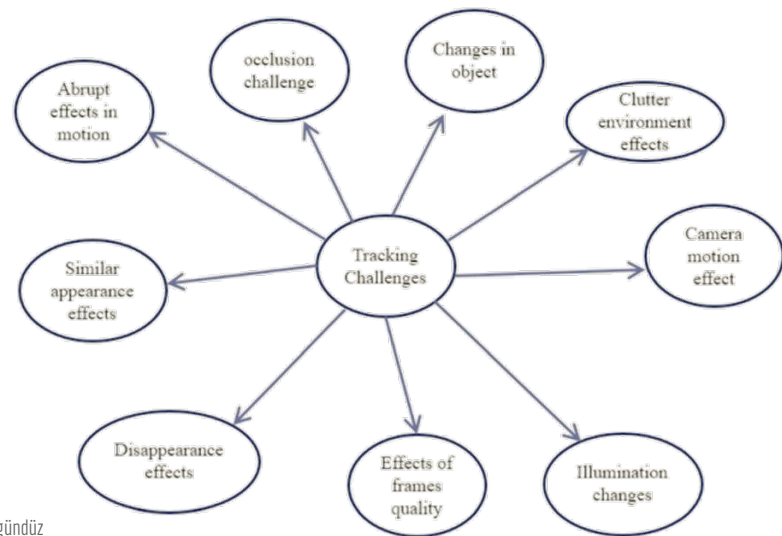
# This week:

- Last week, we talked about the latest achievement in sequence modelling, namely the Transformers.
- This week we are going to study another application of sequence models: Deep Trackers

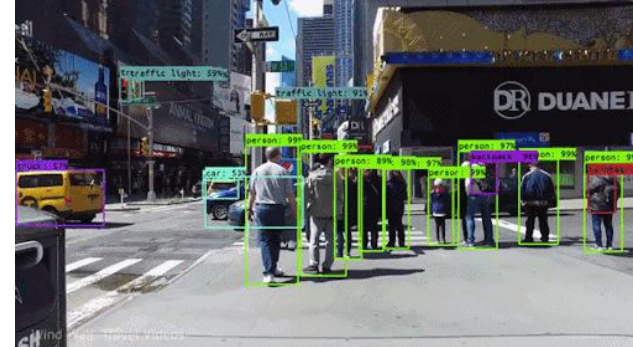


# RNN Trackers

- Tracking is by definition a seq2seq problem. The input is a (for a visual tracker) a set of images (i.e. video) and the output is the pixel location of the tracked target.
- Main challenges are:
  - Occlusion
  - Drifting
  - Association (for multiple targets)
  - Appearance change
  - *others...*



# RNN Trackers



- Tracking is by definition a seq2seq problem. The input is a (for a visual tracker) a set of images (i.e. video) and the output is the pixel location of the tracked target.
- Main challenges are:
  - Occlusion
  - Drifting
  - Association (for multiple targets)
  - Appearance change
  - *others...*



# RNN Trackers

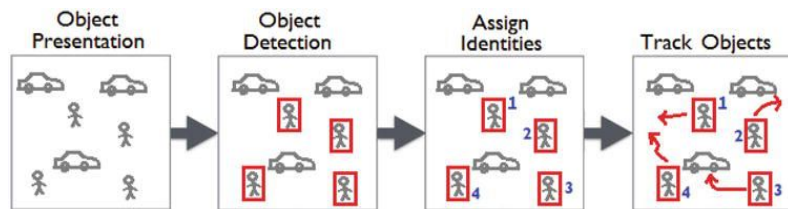
(first what is tracking?)



- Conventionally (before deep learning) visual trackers attacked the problem as
  - Creating an appearance model for the object (and the background)
  - Updating the model with each incoming frame
- The problem had a nature that required a history of the model/signal.
- The models and methods were transparent and hand-crafted

# RNN Trackers

- RNNs with their ability to model sequences have a potential to solve all of the problems that visual trackers face.
- RNN-based trackers estimate the target's bounding box by directly regressing it from the RNN hidden state
- Many CNN+RNN models for tracking have been proposed in the last decade.
- Let's visit some of them.



Yihan Du<sup>1</sup>, Yan Yan<sup>1\*</sup>, Si Chen<sup>2</sup>, Yang Hua<sup>3</sup>, Hanzi Wang<sup>1</sup>

<sup>1</sup>School of Information Science and Engineering, Xiamen University, China

<sup>2</sup>School of Computer and Information Engineering, Xiamen University of Technology, China

<sup>3</sup>School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, UK

Email: yihandu@stu.xmu.edu.cn, {yanyan, hanzi.wang}@xmu.edu.cn, chensi@xmut.edu.cn, y.hua@qub.ac.uk

## RNN-based Trackers

- These guys simply adopt an LSTM to decide where the object to be tracked is using CNN features.
- Thus it is a very simple CNN+RNN solution to single target tracking.
- The LSTMs keep the history of the tracking

*Abstract*—Convolutional Neural Networks (CNNs) have shown outstanding performance in visual object tracking. However, most of classification-based tracking methods using CNNs are time-consuming due to expensive computation of complex online fine-tuning and massive feature extractions. Besides, these methods suffer from the problem of over-fitting since the training and testing stages of CNN models are based on the videos from the same domain. Recently, matching-based tracking methods (such as Siamese networks) have shown remarkable speed superiority, while they cannot well address target appearance variations and complex scenes for inherent lack of online adaptability and background information. In this paper, we propose a novel

performance at the cost of high computational complexity due to massive proposal evaluation and sophisticated online fine-tuning. Besides, some high-accuracy trackers (e.g., MDNet [1] and SANet [5]) use videos from the same domain or two intersecting datasets (e.g., OTB [6] and VOT [7]) to train and test their models, which leads to the problem of over-fitting.

Matching-based tracking methods [3, 4] match the candidate patches with the target template and do not involve any updating procedures. Thus, they can operate at real-time speeds [3, 4]. However, the lack of online adaptability and background

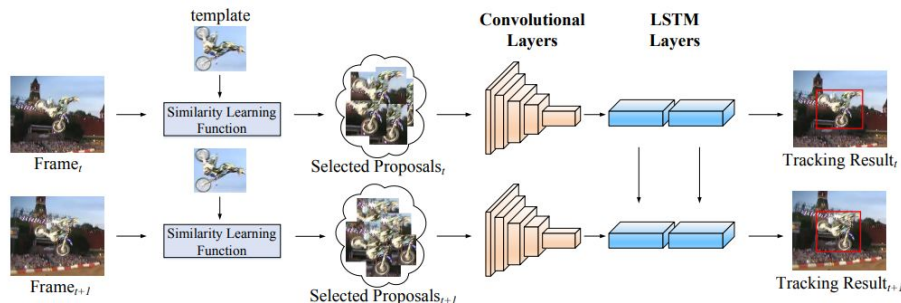


Fig. 1. Pipeline of the proposed method for visual object tracking. The arrows from the LSTM layers of frame<sub>t</sub> to those of frame<sub>t+1</sub> denote the forward propagation in time of memory information.

# RNN-based Trackers

- These guys take it one step ahead and feed the image directly to an RNN model (actually a hybrid model, the convLSTM)
- ConvLSTMs are convolutional layers with hidden states.

Tianyu Yang    Antoni B. Chan  
City University of Hong Kong

tianyyang8-c@my.cityu.edu.hk, abchan@cityu.edu.hk

## Abstract

Recently using convolutional neural networks (CNNs) has gained popularity in visual tracking, due to its robust feature representation of images. Recent methods perform online tracking by fine-tuning a pre-trained CNN model to the specific target object using stochastic gradient descent (SGD) back-propagation, which is usually time-consuming. In this paper, we propose a recurrent filter generation methods for visual tracking. We directly feed the target's image patch to a recurrent neural network (RNN) to estimate an object-specific filter for tracking. As the video sequence is a spatiotemporal data, we extend the matrix multiplications of the fully-connected layers of the RNN to a convolutional

training a discriminative model to classify the target from the background. This self-updating paradigm assumes that the object's appearance changes smoothly, but is inappropriate in challenging situations such as heavy occlusion, illumination changes and abrupt motion. Several methods adopt multiple experts [46], multiple instance learning [2], or short and long term memory stores [19] to address the problem of drastic appearance changes. Recent advances using CNNs for object recognition and detection has inspired tracking algorithms to employ the discriminative features learned by CNNs. In particular, [27, 33, 8] feed the CNN features into a traditional visual tracker, the correlation filter [18], to get a response map for target's estimated location.

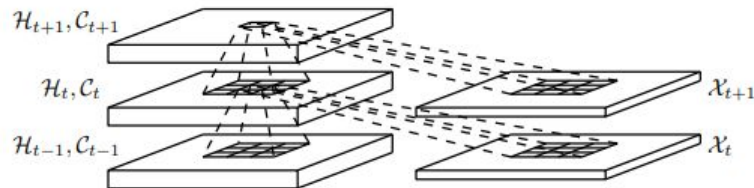


Figure 2: Inner structure of ConvLSTM



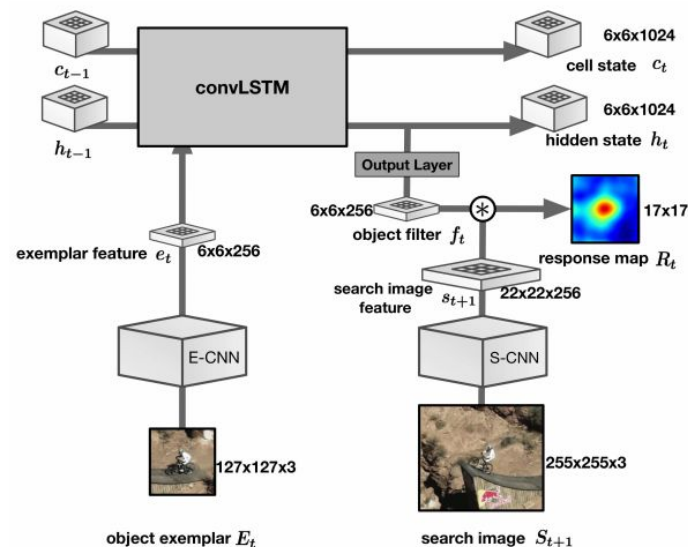
# RNN-based Trackers

- At time step  $t$ , a CNN (E-CNN) extracts features from the exemplar image patch.
- Using the previous hidden and cell states,  $h_{t-1}$  and  $c_{t-1}$ , as well as the current exemplar feature map  $e_t$ , the convolutional LSTM memorizes the appearance information of the target by updating its cell and hidden states.
- The target object filter is generated by passing the new hidden state  $h_t$  through an output convolutional layer.
- A feature map is extracted from the searching image (next frame) using another CNN (S-CNN), which is convolved by the target object filter, resulting in a response map that is used to locate the target.

## Abstract

Recently using convolutional neural networks (CNNs) has gained popularity in visual tracking, due to its robust feature representation of images. Recent methods perform online tracking by fine-tuning a pre-trained CNN model to the specific target object using stochastic gradient descent (SGD) back-propagation, which is usually time-consuming. In this paper, we propose a recurrent filter generation methods for visual tracking. We directly feed the target's image patch to a recurrent neural network (RNN) to estimate an object-specific filter for tracking. As the video sequence is a spatiotemporal data, we extend the matrix multiplications of the fully-connected layers of the RNN to a convolu-

training a discriminative model to classify the target from the background. This self-updating paradigm assumes that the object's appearance changes smoothly, but is inappropriate in challenging situations such as heavy occlusion, illumination changes and abrupt motion. Several methods adopt multiple experts [46], multiple instance learning [2], or short and long term memory stores [19] to address the problem of drastic appearance changes. Recent advances using CNNs for object recognition and detection has inspired tracking algorithms to employ the discriminative features learned by CNNs. In particular, [27, 33, 8] feed the CNN features into a traditional visual tracker, the correlation filter [18], to get a response map for target's estimated location.



# RNN-based Trackers

- What about multiple targets?
- For radar tracking, the input is the azimuth angle and the distance.
- These guys simply preprocess the radar data, then feed it to a BiLSTM.

Full Length Article

DeepMTT: A deep learning maneuvering target-tracking algorithm based on bidirectional LSTM network

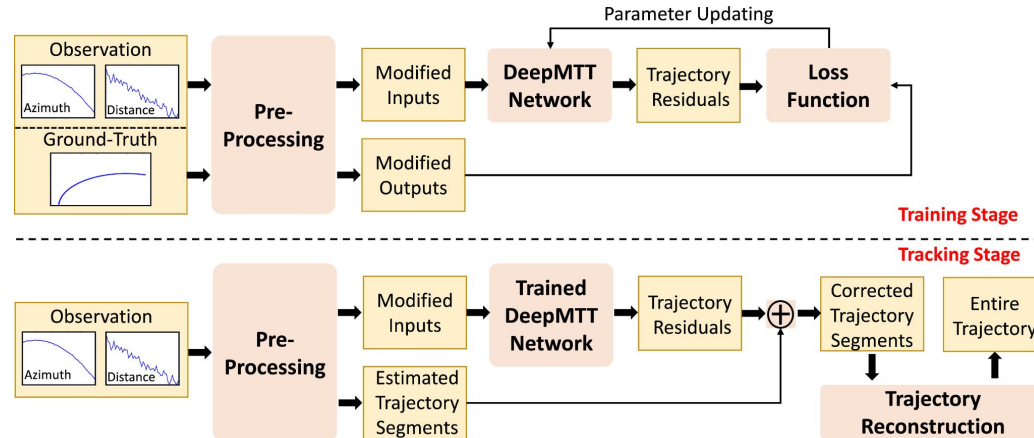
Jingxian Liu <sup>a, b</sup>, Zulin Wang <sup>1, a</sup>, Mai Xu <sup>2, a, c</sup>

[Show more](#)

[+](#) Add to Mendeley [↻](#) Share [🗨](#) Cite

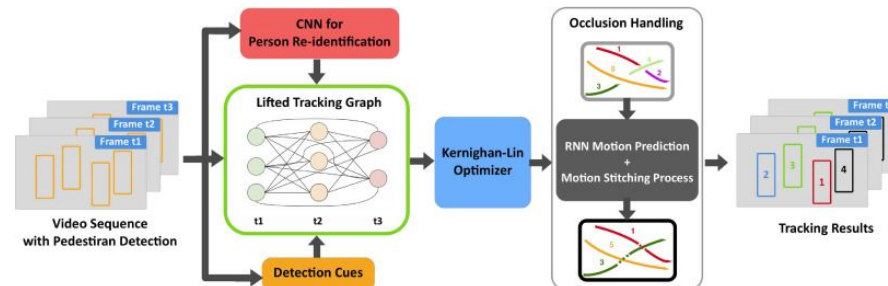
<https://doi.org/10.1016/j.inffus.2019.06.012>

[Get rights and content](#)



# RNN-based Trackers


- For multiple target detection usually tracking-by-detection paradigm is followed.
- These guys propose a dual CNN-RNN model for multiple people tracking.
- They start by detecting targets, then use CNN+RNN to first track then associate targets.



## A dual CNN-RNN for multiple people tracking

 Maryam Babaei , Zimu Li, Gerhard Rigoll

 Show more 

 + Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.neucom.2019.08.008>

Get rights and content

### Abstract

In this paper, we present a deep learning-based approach, namely a dual CNN-RNN for multiple people tracking. We follow tracking-by-detection paradigm by first training a CNN to measure the similarity of two detection boxes. To solve the data association (DA) problem, we build a graph with nodes as detections and edge costs that are the outputs of a CNN. The general minimum cost lifted multi-cut problem (LMP) and corresponding optimization algorithms are utilized to solve the DA problem. To tackle occlusion and ID-switch problems, an RNN network is proposed to predict the nonlinear motion of people. Moreover, we utilize target motion information to stitch tracklets and build long trajectories. The results of our experiments conducted on Multiple Object Tracking Benchmark 2016 (MOT2016) confirm the efficiency of the proposed algorithm.

# RNN-based Trackers

- They model the problem with a seq2seq architecture.

## A dual CNN-RNN for multiple people tracking

Maryam Babaei, Zimu Li, Gerhard Rigoll

Show more

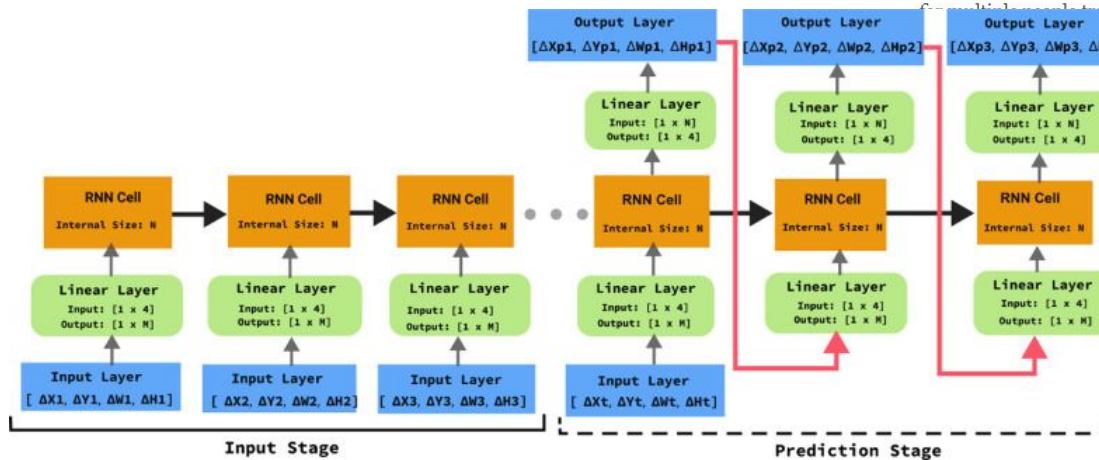
+ Add to Mendeley   Share   Cite

<https://doi.org/10.1016/j.neucom.2019.08.008>

Get rights and content

### Abstract

In this paper, we present a deep learning-based approach, namely a dual CNN-RNN for multiple people tracking. We follow tracking-by-detection paradigm by first detecting objects and then tracking them based on the similarity of two detection boxes. To solve the data association problem, we build a graph with nodes as detections and edge costs between nodes. The general minimum cost lifted multi-cut problem optimization algorithms are utilized to solve the DA and ID-switch problems, an RNN network is proposed to model the motion of people. Moreover, we utilize target motion models and build long trajectories. The results of our Multiple Object Tracking Benchmark 2016 (MOT2016) proposed algorithm.



# Transformer-based Trackers

Xin Chen<sup>1</sup>\*, Bin Yan<sup>1</sup>\*, Jiawen Zhu<sup>1</sup>, Dong Wang<sup>1</sup>†, Xiaoyun Yang<sup>3</sup> and Huchuan Lu<sup>1,2</sup>

<sup>1</sup>School of Information and Communication Engineering, Dalian University of Technology, China

<sup>2</sup>Peng Cheng Laboratory <sup>3</sup>Remark AI

{chenxin3131, yan.bin, jiawen}@mail.dlut.edu.cn

wdice@dlut.edu.cn, xyang@remarkholdings.com, lhchuan@dlut.edu.cn

## Abstract

Correlation acts as a critical role in the tracking field, especially in recent popular Siamese-based trackers. The correlation operation is a simple fusion manner to consider the similarity between the template and the search region. However, the correlation operation itself is a local linear matching process, leading to lose semantic information and fall into local optimum easily, which may be the bottleneck of designing high-accuracy tracking algorithms. Is there any better feature fusion method than correlation? To address this issue, inspired by Transformer, this work presents a novel attention-based feature fusion network, which effectively combines the template and search region features solely using attention. Specifically, the proposed method includes an ego-context augment module based on self-attention and a cross-feature augment module based on cross-attention. Finally, we present a Transformer tracking (named TransT) method based on the Siamese-like feature extraction backbone, the designed attention-based fusion mechanism, and the classification and regression head. Experiments show that our TransT achieves very promis-

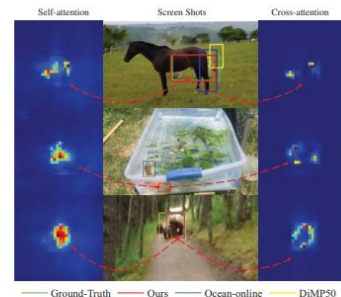


Figure 1. Tracking results of TransT and two state-of-the-art trackers. Our tracker is more robust and accurate in handling various challenges, such as occlusion, similar object interference, motion blur.

- Correlation acts as a critical role in the tracking field, especially in recent popular Siamese-based trackers.
- The correlation operation is a simple fusion manner to consider the similarity between the template and the search region.
- However, the correlation operation itself is a local linear matching process, leading to lose semantic information and fall into local optimum easily, which may be the bottleneck of designing high-accuracy tracking algorithms.
- Is there any better feature fusion method than correlation?



# Transformer-based Trackers

- Is there any better feature fusion method than correlation?
- To address this issue, inspired by Transformer, this work presents a novel attention-based feature fusion network, which effectively combines the template and search region features solely using attention.
- Specifically, the proposed method includes an ego-context augment module based on self-attention and a cross-feature augment module based on cross-attention.

Xin Chen<sup>1</sup>\*, Bin Yan<sup>1</sup>\*, Jiawen Zhu<sup>1</sup>, Dong Wang<sup>1</sup>†, Xiaoyun Yang<sup>3</sup> and Huchuan Lu<sup>1,2</sup>

<sup>1</sup>School of Information and Communication Engineering, Dalian University of Technology, China

<sup>2</sup>Peng Cheng Laboratory <sup>3</sup>Remark AI

{chenxin3131, yan.bin, jiawen}@mail.dlut.edu.cn

wdice@dlut.edu.cn, xyang@remarkholdings.com, lhchuan@dlut.edu.cn

## Abstract

Correlation acts as a critical role in the tracking field, especially in recent popular Siamese-based trackers. The correlation operation is a simple fusion manner to consider the similarity between the template and the search region. However, the correlation operation itself is a local linear matching process, leading to lose semantic information and fall into local optimum easily, which may be the bottleneck of designing high-accuracy tracking algorithms. Is there any better feature fusion method than correlation? To address this issue, inspired by Transformer, this work presents a novel attention-based feature fusion network, which effectively combines the template and search region features solely using attention. Specifically, the proposed method includes an ego-context augment module based on self-attention and a cross-feature augment module based on cross-attention. Finally, we present a Transformer tracking (named TransT) method based on the Siamese-like feature extraction backbone, the designed attention-based fusion mechanism, and the classification and regression head. Experiments show that our TransT achieves very promis-

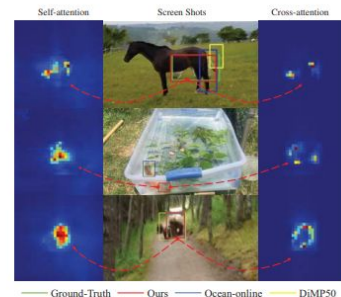


Figure 1. Tracking results of TransT and two state-of-the-art trackers. Our tracker is more robust and accurate in handling various challenges, such as occlusion, similar object interference, motion blur.

# Transformer-based Trackers

- A new tracking architecture with an encoder-decoder transformer as the key component. The encoder models the global spatio-temporal feature dependencies between target objects and search regions, while the decoder learns a query embedding to predict the spatial positions of the target objects. Our method casts object tracking as a direct bounding box prediction problem, without using any proposals or predefined anchors. With the encoder-decoder transformer, the prediction of objects just uses a simple fully-convolutional network, which estimates the corners of objects directly. The whole method is end-to-end, does not need any postprocessing steps such as cosine window and bounding box smoothing, thus largely simplifying existing tracking pipelines. The proposed tracker achieves state-of-the-art performance on multiple challenging short-term and long-term benchmarks,
- The encoder models the global spatio-temporal feature dependencies between target objects and search regions,
- while the decoder learns a query embedding to predict the spatial positions of the target objects.

## Abstract

In this paper, we present a new tracking architecture with an encoder-decoder transformer as the key component. The encoder models the global spatio-temporal feature dependencies between target objects and search regions, while the decoder learns a query embedding to predict the spatial positions of the target objects. Our method casts object tracking as a direct bounding box prediction problem, without using any proposals or predefined anchors. With the encoder-decoder transformer, the prediction of objects just uses a simple fully-convolutional network, which estimates the corners of objects directly. The whole method is end-to-end, does not need any postprocessing steps such as cosine window and bounding box smoothing, thus largely simplifying existing tracking pipelines. The proposed tracker achieves state-of-the-art performance on multiple challenging short-term and long-term benchmarks,

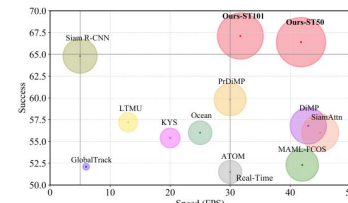


Figure 1: Comparison with state-of-the-arts on LaSOT [15]. We visualize the Success performance with respect to the Frames-Per-Seconds (fps) tracking speed. The circle size indicates a weighted sum of the tracker's speed (x-axis) and success score (y-axis). The larger, the better. Ours-ST101 and Ours-ST50 indicate the proposed trackers with ResNet-101 and ResNet-50 as backbones, respectively. Better viewed in color.

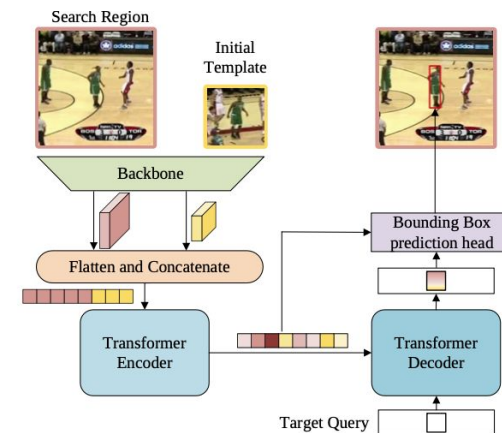


Figure 2: Framework for spatial-only tracking.

# Additional Reading & References

- [Object-Adaptive LSTM Network for Visual Tracking](#)
- [Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting](#)
- [Recurrent Filter Learning for Visual Tracking](#)
- [A dual CNN-RNN for multiple people tracking](#)
- [DeepMTT: A deep learning maneuvering target-tracking algorithm based on bidirectional LSTM network](#)
- [https://openaccess.thecvf.com/content/CVPR2021/papers/Chen\\_Transformer\\_Tracking\\_CVPR\\_2021\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2021/papers/Chen_Transformer_Tracking_CVPR_2021_paper.pdf)
- <https://www.robots.ox.ac.uk/~luca/siamese-fc.html>
- [https://openaccess.thecvf.com/content/ICCV2021/papers/Yan\\_Learning\\_Spatio-Temporal\\_Transformer\\_for\\_Visual\\_Tracking\\_ICCV\\_2021\\_paper.pdf](https://openaccess.thecvf.com/content/ICCV2021/papers/Yan_Learning_Spatio-Temporal_Transformer_for_Visual_Tracking_ICCV_2021_paper.pdf)