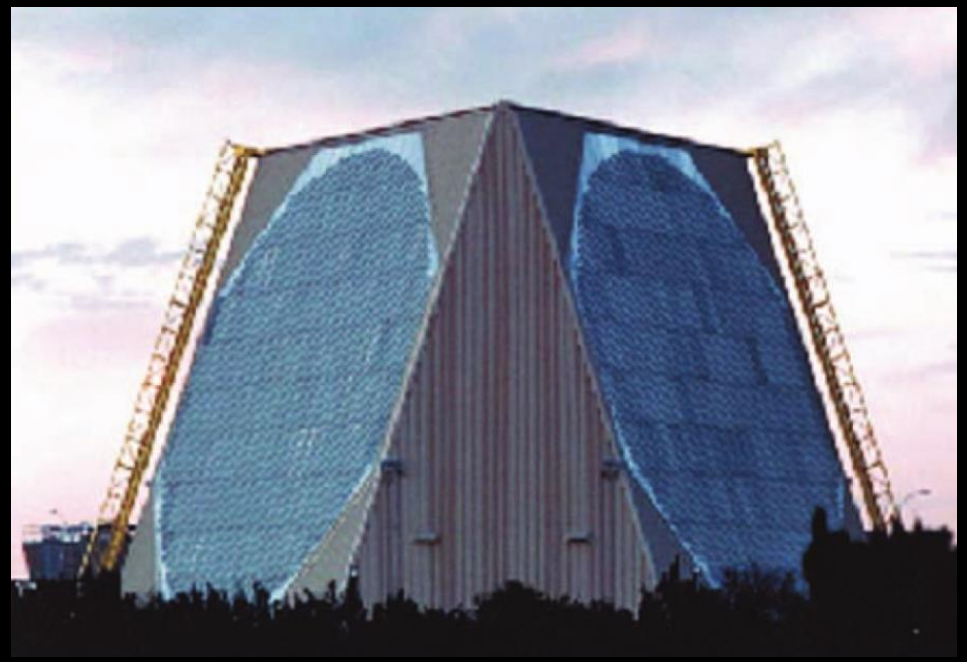
A Survey on Leveraging Deep Neural Networks for Object Tracking

Sebastian Krebs, Bharanidhar Duraisamy, and Fabian Flohr
Daimler AG, Research and Development, Ulm (Germany)
Contact: Sebastian.Krebs@Daimler.com

Mercedes-Benz
The best or nothing.

Tracking
- General

- Originated from aerospace applications in the 1960s
- Estimating the state of one or several targets over time
- Based on noisy measurements from one or multiple sensors



From: Y. Bar-Shalom et al.
„The Probabilistic Data Association Filter“,
in IEEE Control Systems, 2009

Tracking
- Autonomous Driving Applications

Motivation

- Robustify detections results
- Extract non-directly observables (velocities)
- Provide information for higher-level systems

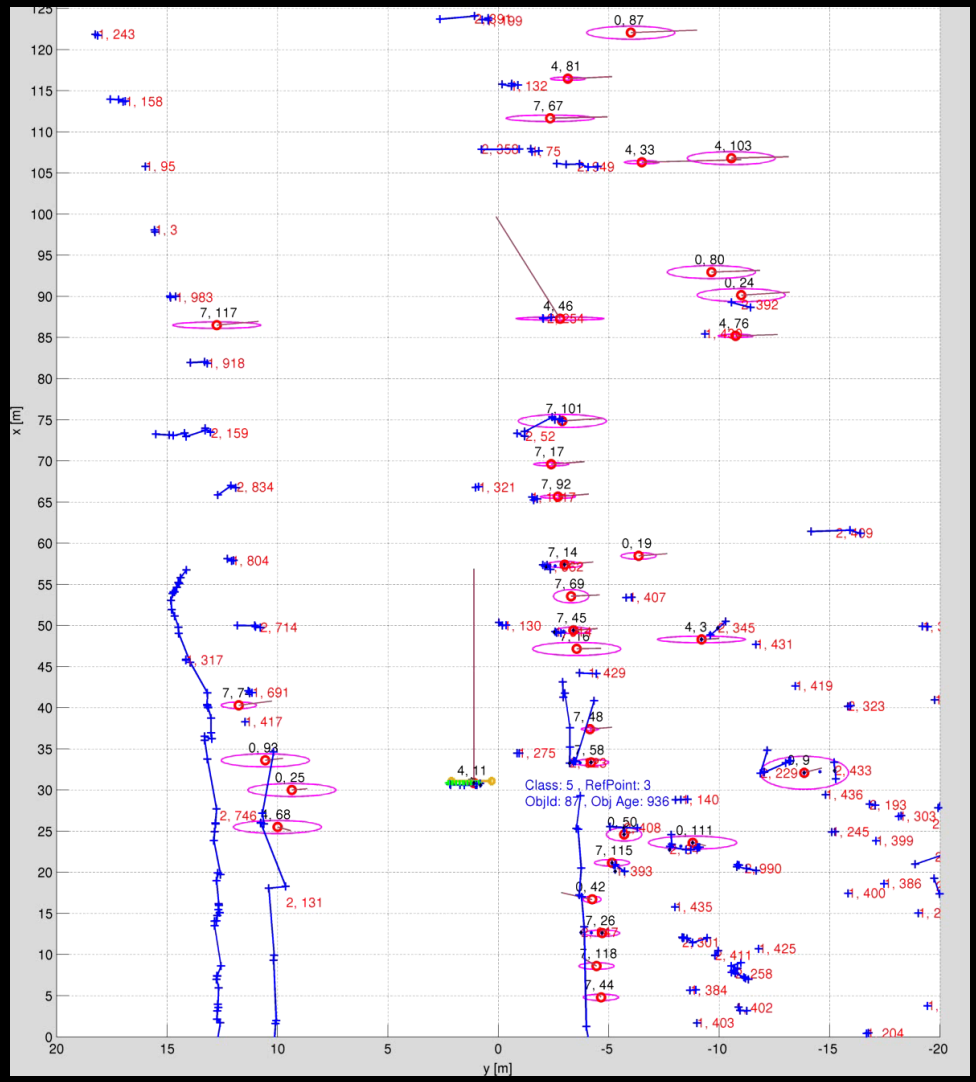
Tracking
- Autonomous Driving Applications

Motivation

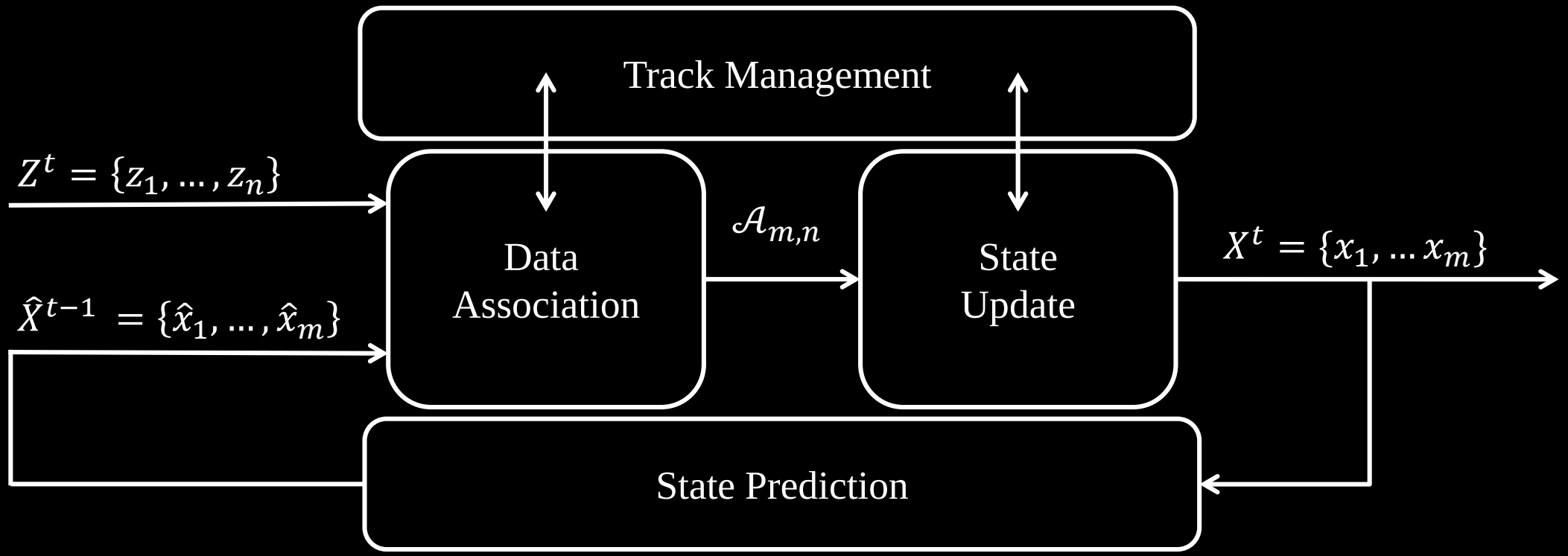
- Robustify detections results
- Extract non-directly observables (velocities)
- Provide information for higher-level systems

Challenges

- Possible high amount of objects
- High proximity of objects
- Agile motion patterns

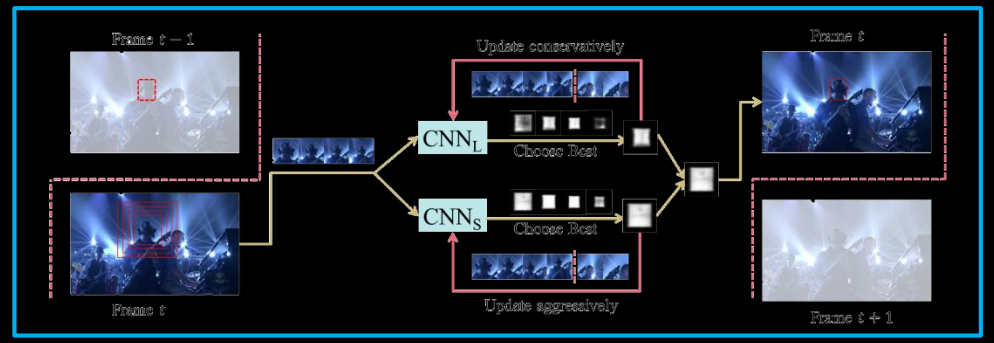


Tracking
- Traditional Object Tracking

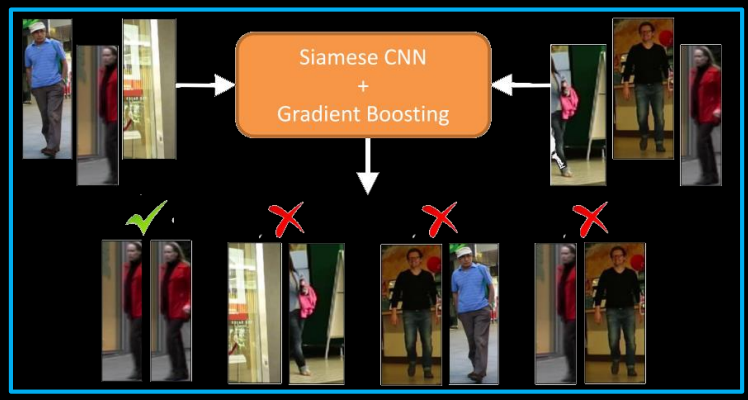


Deep Learning for Object Tracking - Overview

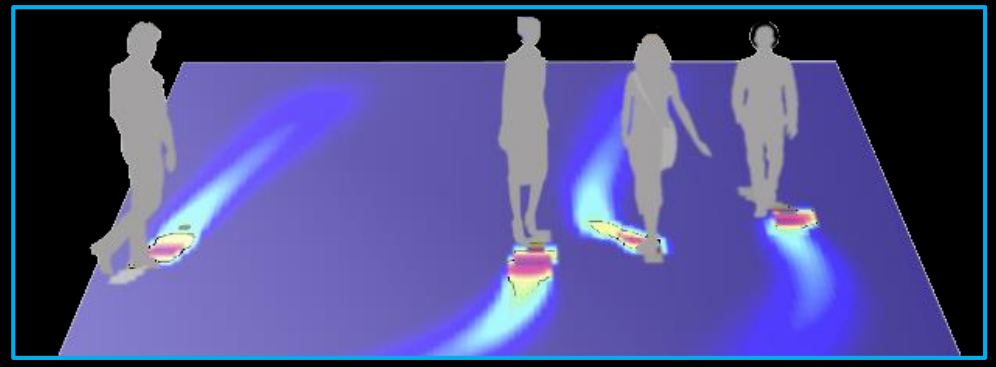
Features



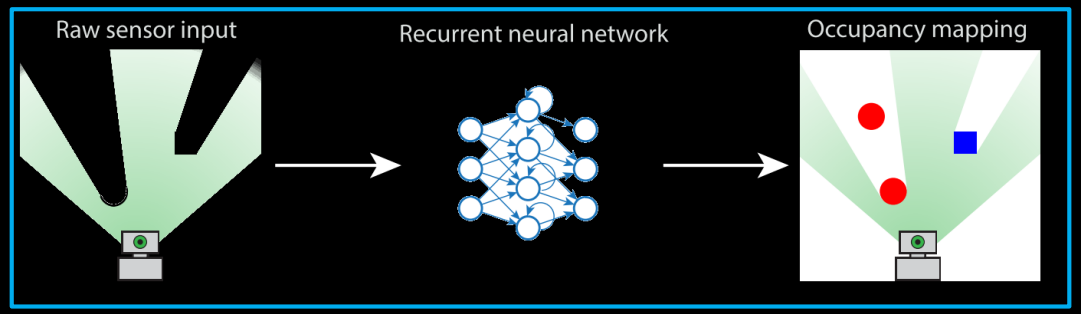
Data Association



Prediction

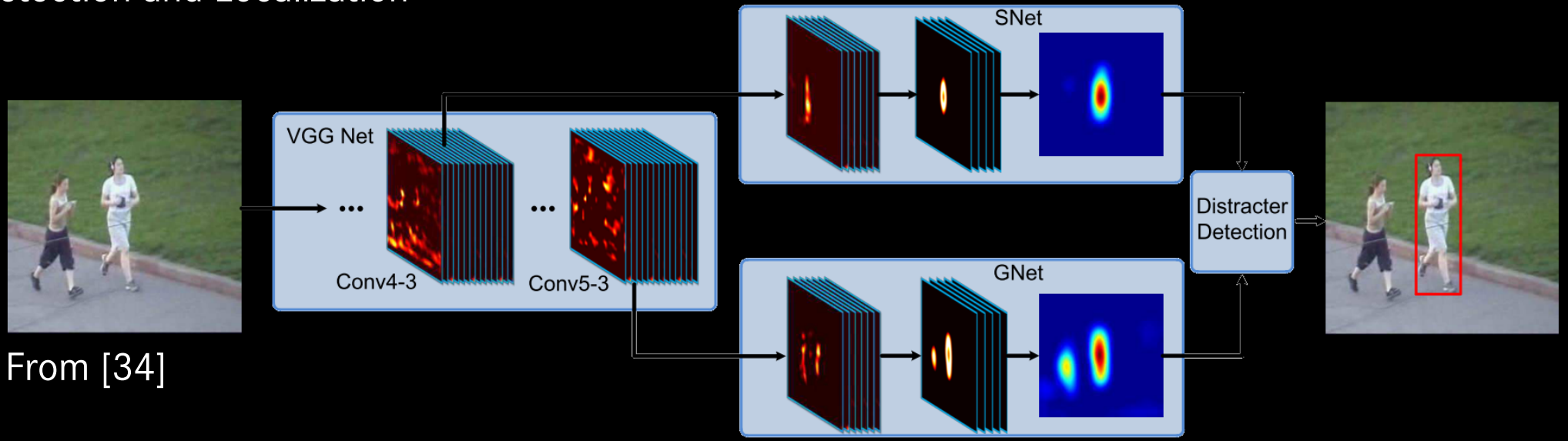


End-to-End



Deep Learning for Object Tracking - Features

- Pre-train network on big image database
- Utilize feature maps from pre-trained network
  - Create and update a model of the tracked object
  - Detection and Localization



From [34]

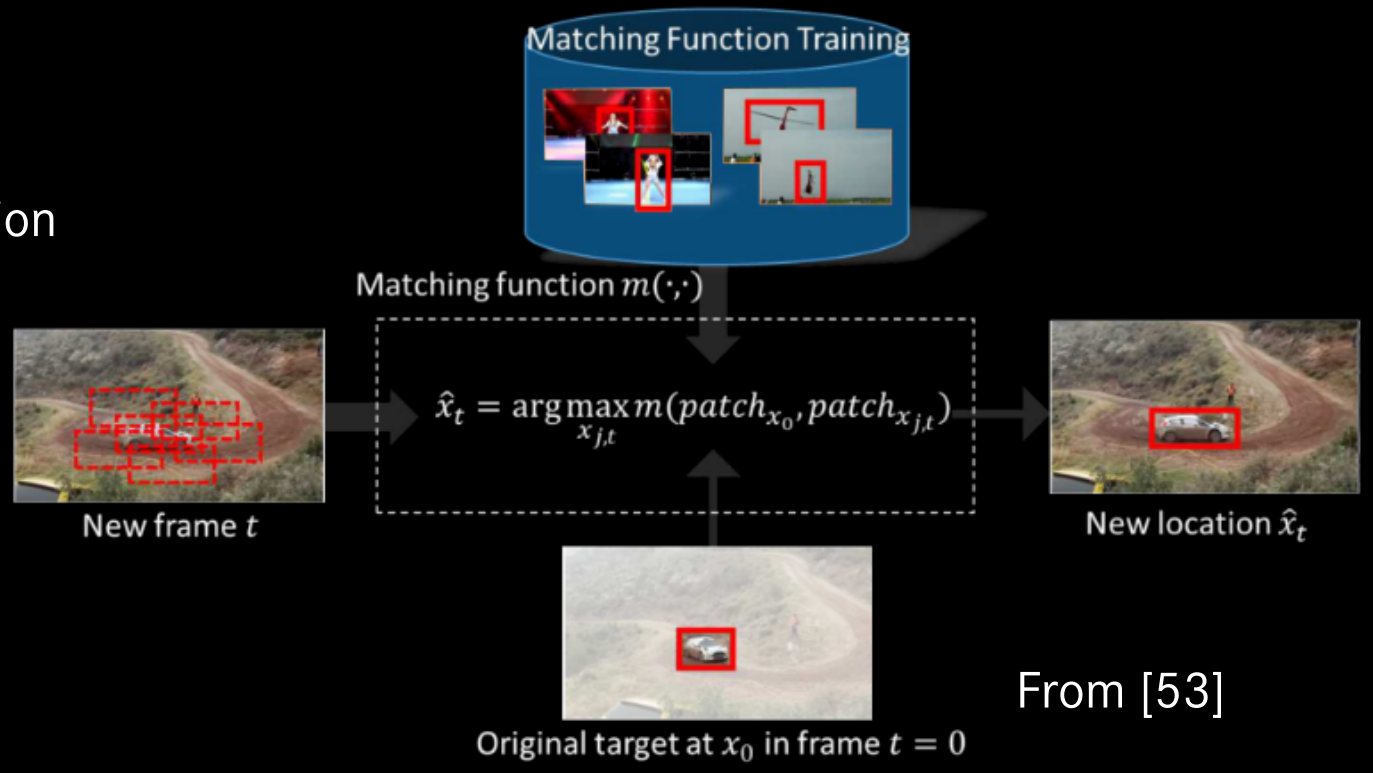
[34] L. Wang, W. Ouyang, X. Wang, and H. Lu, “Visual Tracking with Fully Convolutional Networks,” in ICCV, 2015

Deep Learning for Object Tracking - Features

Method Name	Network	Features	Integration Remarks
DLT [17]	Stacked Denoising Autoencoder (SDAE)	Pre-trained encoder with classification layer	Network output is used as conficence for a particle filter based tracking approach.
SO-DLT [33]	Structured Output CNN	50x50 Probability Map	During tracking two CNNs are fine-tuned on the desired target.
Wang et al. [34]	VGG	conv4-3, conv5-3	Feature map selection, two networks for generic and specific features, distractor removal
Chi et al. [35]	VGG, Dual Network	conv4-3, conv5-3, boundary maps	Dual network is trained and updated to fine tune features for a specific target.
Ma et al. [36]	VGG	conv3-4, conv4-4, conv5-4	Learn adaptive linear correlation filter per layer to obtain response maps, to infer target location
Hong et al. [52]	R-CNN	Outputs from fc 1	R-CNN features are classified by an online-learned SVM, back propagated trough network to obtain saliency maps. Bayesian filtering performed on combined saliency maps

Deep Learning for Object Tracking - Data Association

- Learn generic similarity measure directly from the data
  - Using Siamese Networks
    Two-stream networks, with shared weight
    Learned with a contrasive loss
- Use of similarity measure during data association



From [53]

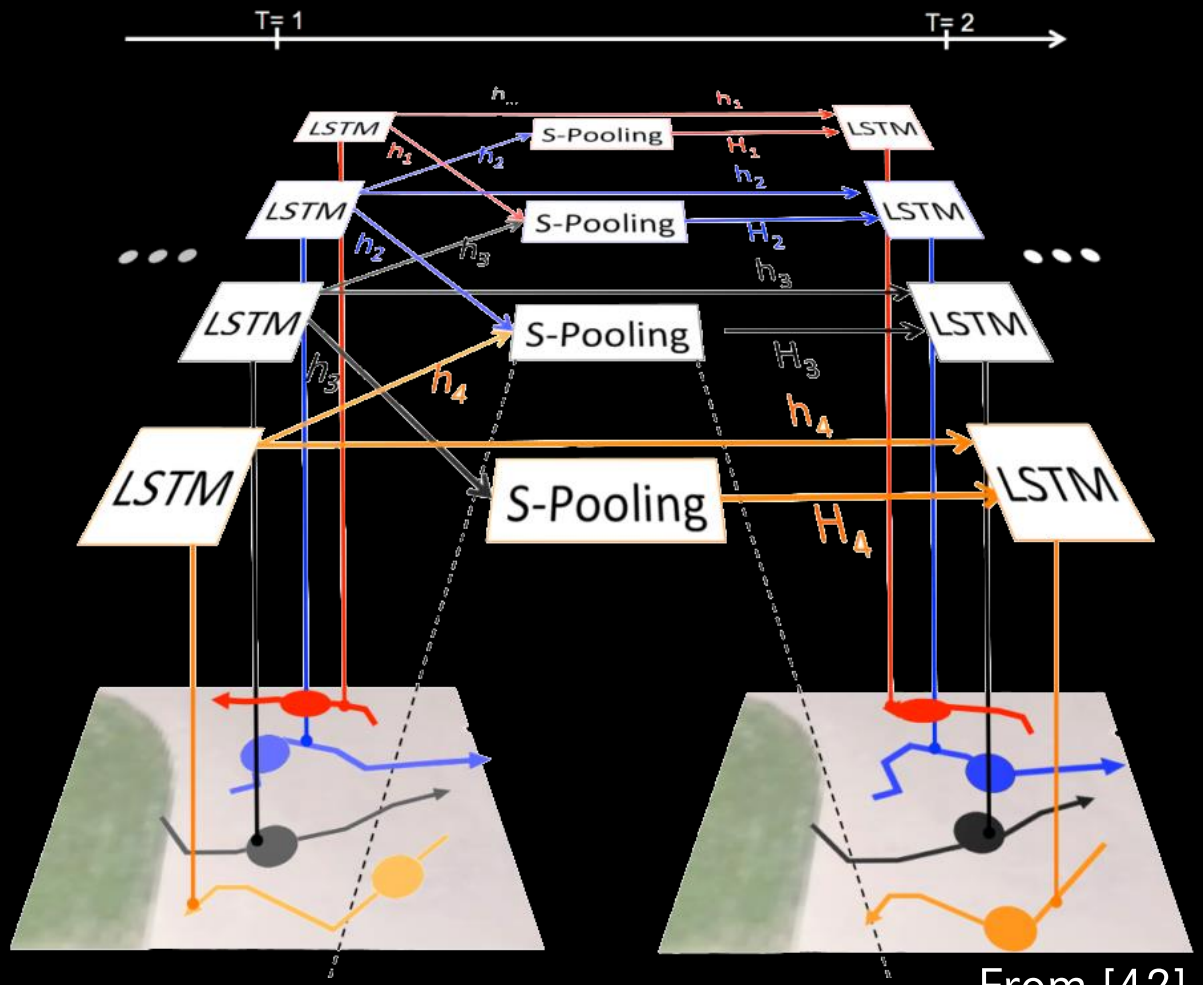
[53] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, "Learning by Tracking: Siamese CNN for Robust target association," in CVPRW, 2016

Deep Learning for Object Tracking - Data Association

Method Name	Similarity Between	Input	Integration Remark
SINT [37]	Target template and candidate boxes	Image patches (pixel values)	Radius sampling to generate candidate patches, similarity measure per proposal box. Box with highest similarity is considered new target position
Leal-Taixe et al. [53]	Detection at time t and t+1	Pixel Values, Optical Flow, Contextual Information	Similarity of flow and pixel patches is calculated by the Siamese network, combined with contextual features to calculate probability of matching. Which is used by the final linear programming tracker.
Varior et al. [38]	Pair of target patches	Local Maximal Occurence (LOMO), Color Names (CN)	Divide patches into horizontal rows, which are interpreted as a sequence.

[37] R. Tao, E. Gavves, and A. W. M. Smeulders, “Siamese Instance Search for Tracking” in CVPR, 2016
[53] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, “Learning by Tracking: Siamese CNN for Robust target association,” in CVPRW, 2016
[38] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, “A Siamese Long Short-Term Memory Architecture for Human Re-Identification” in ECCV, 2016

Deep Learning for Object Tracking - Prediction



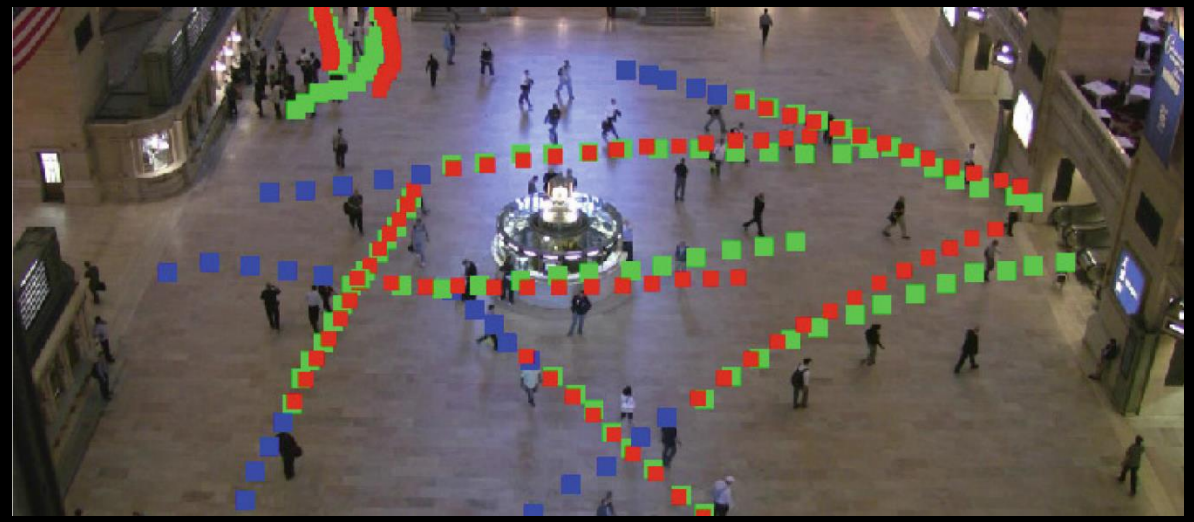
From [42]

Social-LSTM [42]

- Predict path of multiple persons
- Each trajectory is predicted by a LSTM using a pre-processed trajectory history
- Inter-object dependencies are captures by social-pooling layers

[42] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human Trajectory Prediction in Crowded Spaces,” in CVPR, 2016

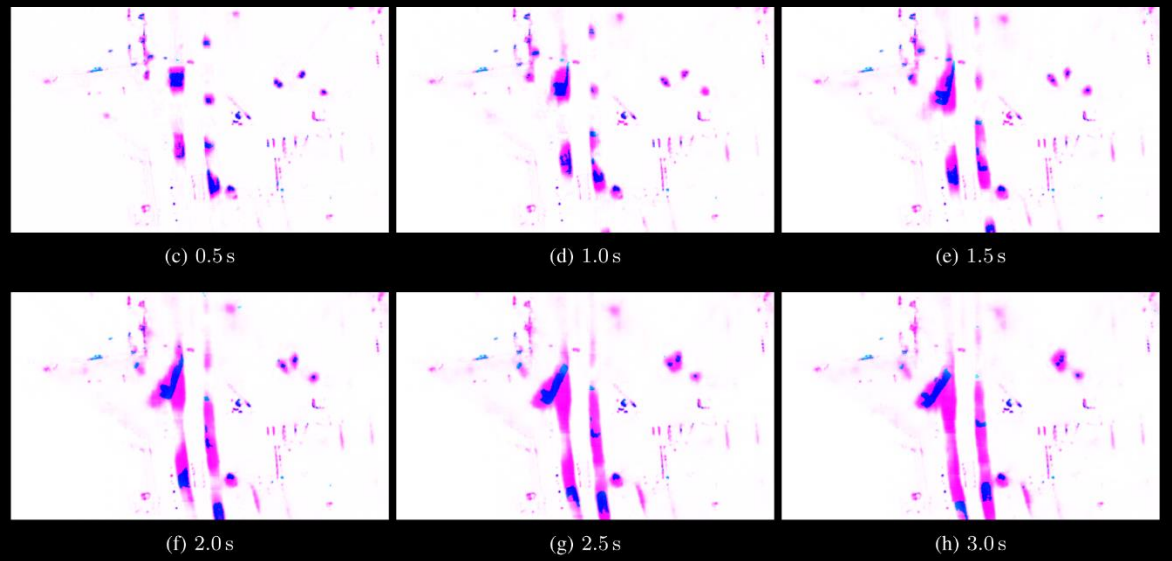
Deep Learning for Object Tracking - Prediction



Behavior-CNN [43]

- Image from a static surveillance camera
- Learn kinematic properties of pedestrians
- Predicts future trajectories based on previous

[43] S. Yi, H. Li, and X. Wang, “Pedestrian Behavior Understanding and Prediction with Deep Neural Networks” in ECCV, 2016



Hoermann et al. [44]

- Dynamic Occupancy Grid Map (DOGMa) as input
- Prediction of whole DOGMa

[44] S. Hoermann, M. Bach, and K. Dietmayer, “Dynamic Occupancy Grid Prediction for Urban Autonomous Driving: A Deep Learning Approach with Fully Automatic Labeling ” in IV, 2017

Deep Learning for Object Tracking - End-to-End

Method Name	Input	Trained on	Network	Integration Remark
Gan et al. [45]	Image, First target bounding box	Artifical data (generic background, shapes)	RCNN (GRU)	Outputs target bounding box. No online fine-tuning. Anonymous tracking.
GOTURN [46]	Current search region, cropped target template	Adjacent video frames and modified images	Two-stream CNN	Outputs target bounding box by regression. No online fine-tuning. Anonymous tracking.
MDNet [19]	Target candidates, initial target position	Real-world videos	Multi-Domain Network	During tracking domain-specific layers are removed. Network fine-tuned during tracking (new classification layer).
ROLO [48]	Raw video frame	ImageNet, Detection (YOLO), videos (LSTM)	YOLO + LSTM	Feature maps of last conv layer and detections results of YOLO are used as input for LSTM. Outputs target bounding box or heat maps.

[45] Q. Gan, Q. Guo, Z. Zhang, and K. Cho, “First Step toward Model-Free, Anonymous Object Tracking with Recurrent Neural Networks” arXiv, 2015
[46] D. Held, S. Thrun, and S. Savarese, “Learning to Track at 100 FPS with Deep Regression Networks” in ECCV, 2016
[19] H. Nam and B. Han, “Learning Multi-domain Convolutional Neural Networks for Visual Tracking” in CVPR, 2016
[48] G. Ning, Z. Zhang, C. Huang, Z. He, X. Ren, H. Wang, “Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking”, arXiv, 2016

Deep Learning for Object Tracking - End-to-End

Method Name	Input	Trained on	Network	Integration Remark
Gan et al. [45]	Image, First target bounding box	Artifical data (generic background, shapes)	RCNN (GRU)	Outputs target bounding box. No online fine-tuning. Anonymous tracking.
GOTURN [46]	Current search region, cropped target template	Adjacent video frames and modified images	Two-stream CNN	Outputs target bounding box by regression. No online fine-tuning. Anonymous tracking.
MDNet [19]	Target candidates, initial target position	Real-world videos	Multi-Domain Network	During tracking domain-specific layers are removed. Network fine-tuned during tracking (new classification layer).
ROLO [48]	Raw video frame	ImageNet, Detection (YOLO), videos (LSTM)	YOLO + LSTM	Feature maps of last conv layer and detections results of YOLO are used as input for LSTM. Outputs target bounding box or heat maps.

Single-object tracking methods without kinematic information

Deep Learning for Object Tracking - End-to-End

DeepTracking [47]

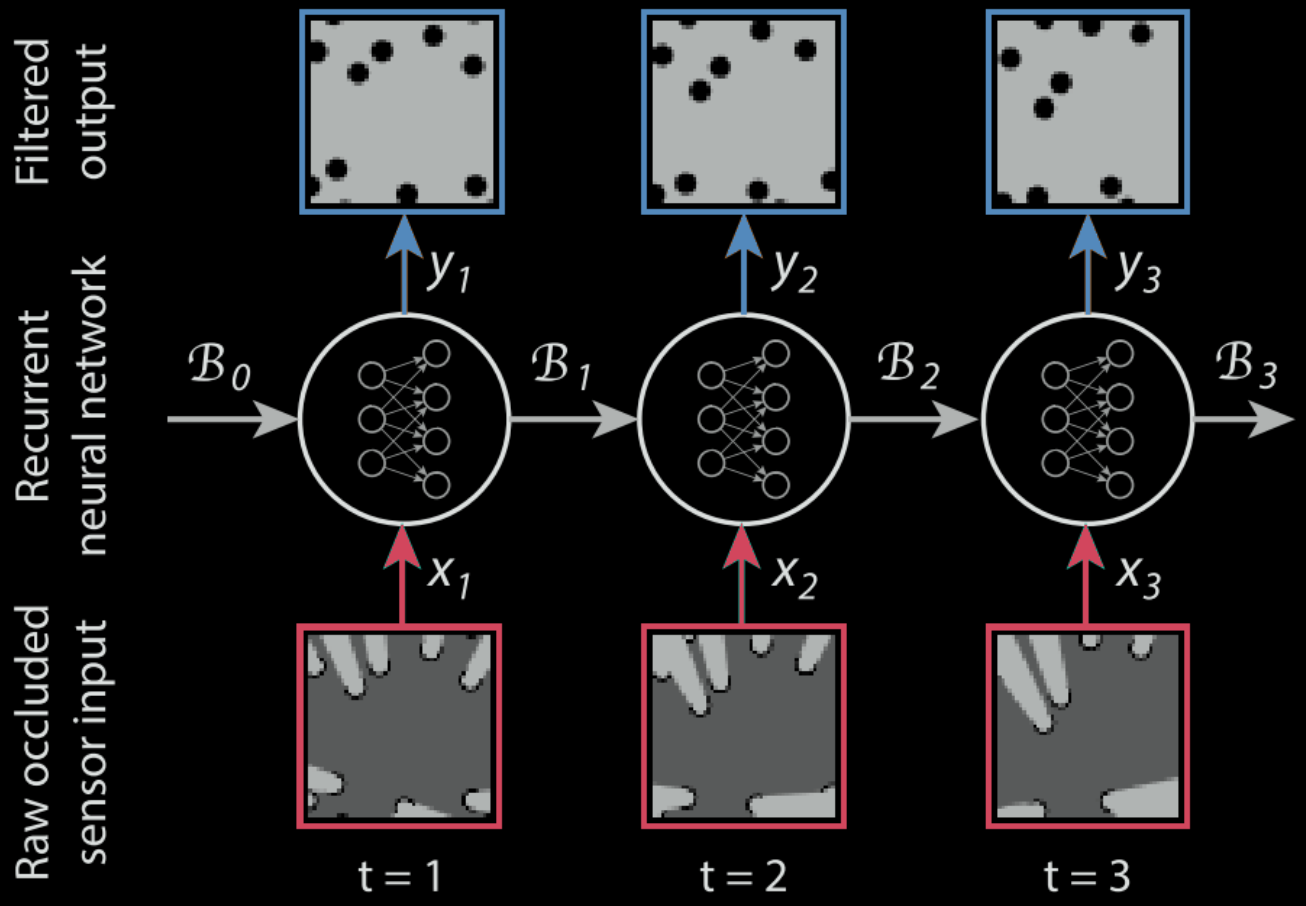
- Raw input from laser scanner
- Predict unoccluded state of the world
- Recurrent Network (GRUs) employed
- Artifical training data

Extension: Ondruska et al. [55]

- Allow classification (object-level)
- Real-data from a traffic intersection

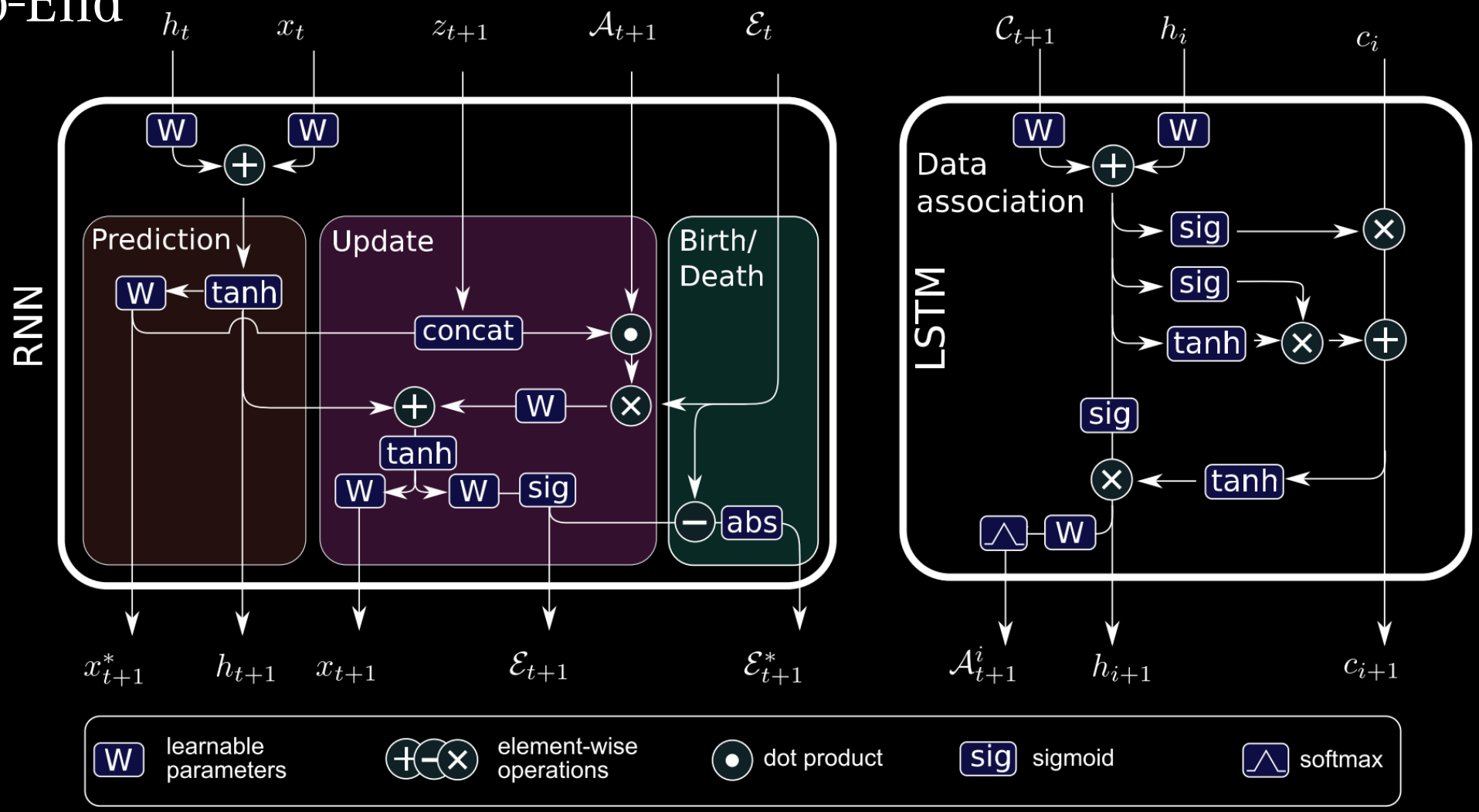
Extension: Dequaire et al. [56]

- Introduces Spatial Transformer Module (STM)
- Applied in a moving vehicle



[47] I. Posner and P. Ondruska, “Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks” in AAAI, 2016

Deep Learning for Object Tracking - End-to-End



[49] A. Milan, S. H. Rezatofighi, A. Dick, K. Schindler, and I. Reid, “Online Multi-target Tracking using Recurrent Neural Networks” in AAAI, 2017

Conclusion

- Most deep-based tracking approaches are tailored by the vision-based detection and classification tasks
- Recurrent Neural Networks are suitable to capture spatio-temporal dependencies
- Most methods lack the explicit modeling of the kinematic state of the target
- Integration of non-image sensor measurements or from multiple sensors still challenging
- Compared to classical deep-based tasks like classification and detection tracking is a “new” research field

Thank you for your attention!

References

[17] N. Wang and D.-Y. Yeung, “Learning a Deep Compact Image Representation for Visual Tracking,” in NIPS, 2013
[19] H. Nam and B. Han, “Learning Multi-domain Convolutional Neural Networks for Visual Tracking” in CVPR, 2016
[33] N. Wang, S. Li, A. Gupta, and D.-Y. Yeung, “Transferring Rich Feature Hierarchies for Robust Visual Tracking” arXiv, 2015
[34] L. Wang, W. Ouyang, X. Wang, and H. Lu, “Visual Tracking with Fully Convolutional Networks,” in ICCV, 2015
[35] Z. Chi, H. Li, H. Lu, and M.-H. Yang, “Dual Deep Network for Visual Tracking” in IEEE Transactions on Image Processing, 2017
[36] C. Ma, J. B. Huang, X. Yang, and M. H. Yang, “Hierarchical Convolutional Features for Visual Tracking” in ICCV 2016
[37] R. Tao, E. Gavves, and A. W. M. Smeulders, “Siamese Instance Search for Tracking” in CVPR, 2016
[38] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, “A Siamese Long Short-Term Memory Architecture for Human Re-Identification” in ECCV, 2016
[42] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human Trajectory Prediction in Crowded Spaces,” in CVPR, 2016
[43] S. Yi, H. Li, and X. Wang, “Pedestrian Behavior Understanding and Prediction with Deep Neural Networks” in ECCV, 2016
[44] S. Hoermann, M. Bach, and K. Dietmayer, “Dynamic Occupancy Grid Prediction for Urban Autonomous Driving: A Deep Learning Approach with Fully Automatic Labeling ” in IV, 2017
[45] Q. Gan, Q. Guo, Z. Zhang, and K. Cho, “First Step toward Model-Free, Anonymous Object Tracking with Recurrent Neural Networks” arXiv, 2015
[46] D. Held, S. Thrun, and S. Savarese, “Learning to Track at 100 FPS with Deep Regression Networks” in ECCV, 2016
[47] I. Posner and P. Ondruska, “Deep Tracking: Seeing Beyond Seeing Using Recurrent Neural Networks” in AAAI, 2016
[48] G. Ning, Z. Zhang, C. Huang, Z. He, X. Ren, H. Wang, “Spatially Supervised Recurrent Convolutional Neural Networks for Visual Object Tracking”, arXiv, 2016
[49] A. Milan, S. H. Rezatofighi, A. Dick, K. Schindler, and I. Reid, “Online Multi-target Tracking using Recurrent Neural Networks” in AAAI, 2017
[53] L. Leal-Taixe, C. Canton-Ferrer, and K. Schindler, “Learning by Tracking: Siamese CNN for Robust target association,” in CVPRW, 2016
[55] P. Ondruska, J. Dequaire, D. Z. Wang, and I. Posner, “End-to-End Tracking and Semantic Segmentation Using Recurrent Neural Networks” arXiv, 2016
[56] J. Dequaire, D. Rao, P. Ondruska, D. Wang, and I. Posner, “Deep Tracking on the Move: Learning to Track the World from a Moving Vehicle using Recurrent Neural Networks” arXiv , 2016