Looking into the Future: Forecasting Quantities with Deep Learning

Lorenzo Seidenari

University of Florence

ARTIFICIAL INTELLIGENCE
DOCTORAL ACADEMY

Hello!

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Lorenzo Seidenari, Associate Professor

lorenzo.seidenari@unifi.it

www.micc.unifi.it/seidenari
www.small-pixels.com

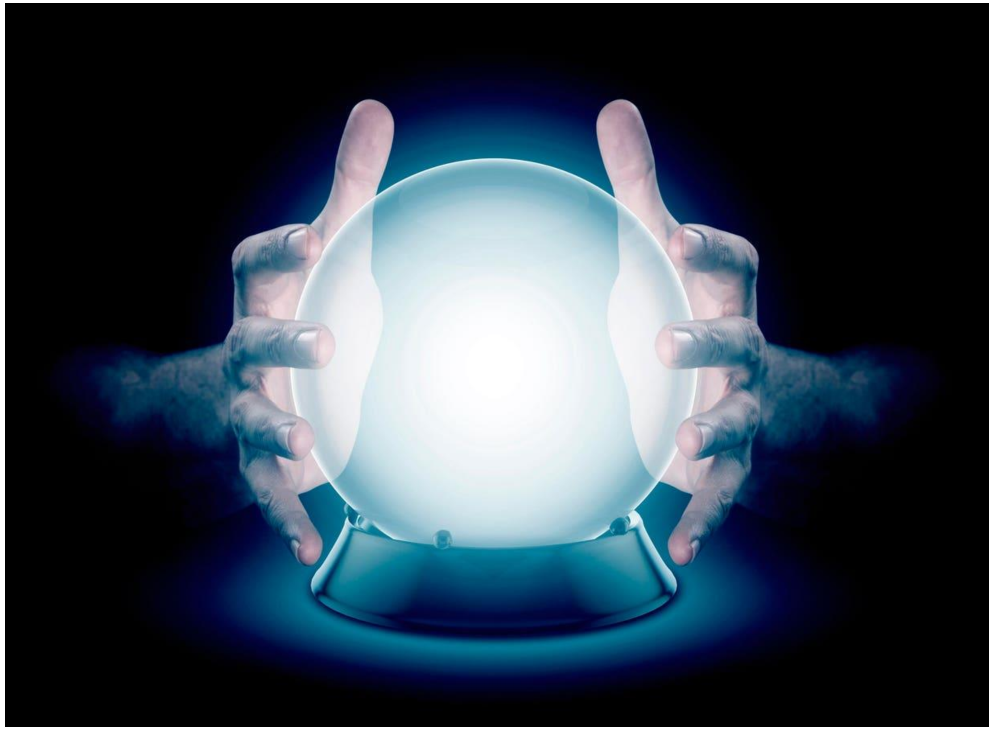
smallpixels

ellis
European Laboratory for Learning and Intelligent Systems

MICC
Media Integration and Communication Center

Outline

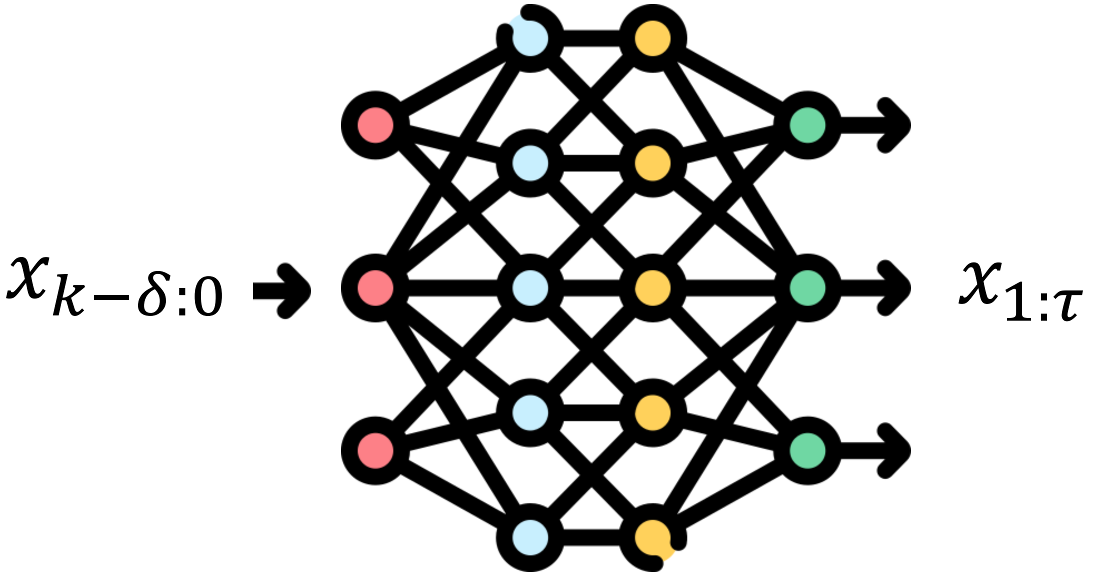
- Preliminaries
    1. Models
    2. Sensing
- Feature forecasting (EGO)
    1. Forecasting Depth and Flow
    2. Behavior Forecasting
- Memory based Trajectory forecasting (BEV)
    1. Forecasting w/o social context
    2. Socially-Aware forecasting
- Foundation models for Time Series
    1. Zero Shot Forecasting with LLMs
    2. Specialized Foundation Models



“Can you tell me what the future holds for me?”

Outline

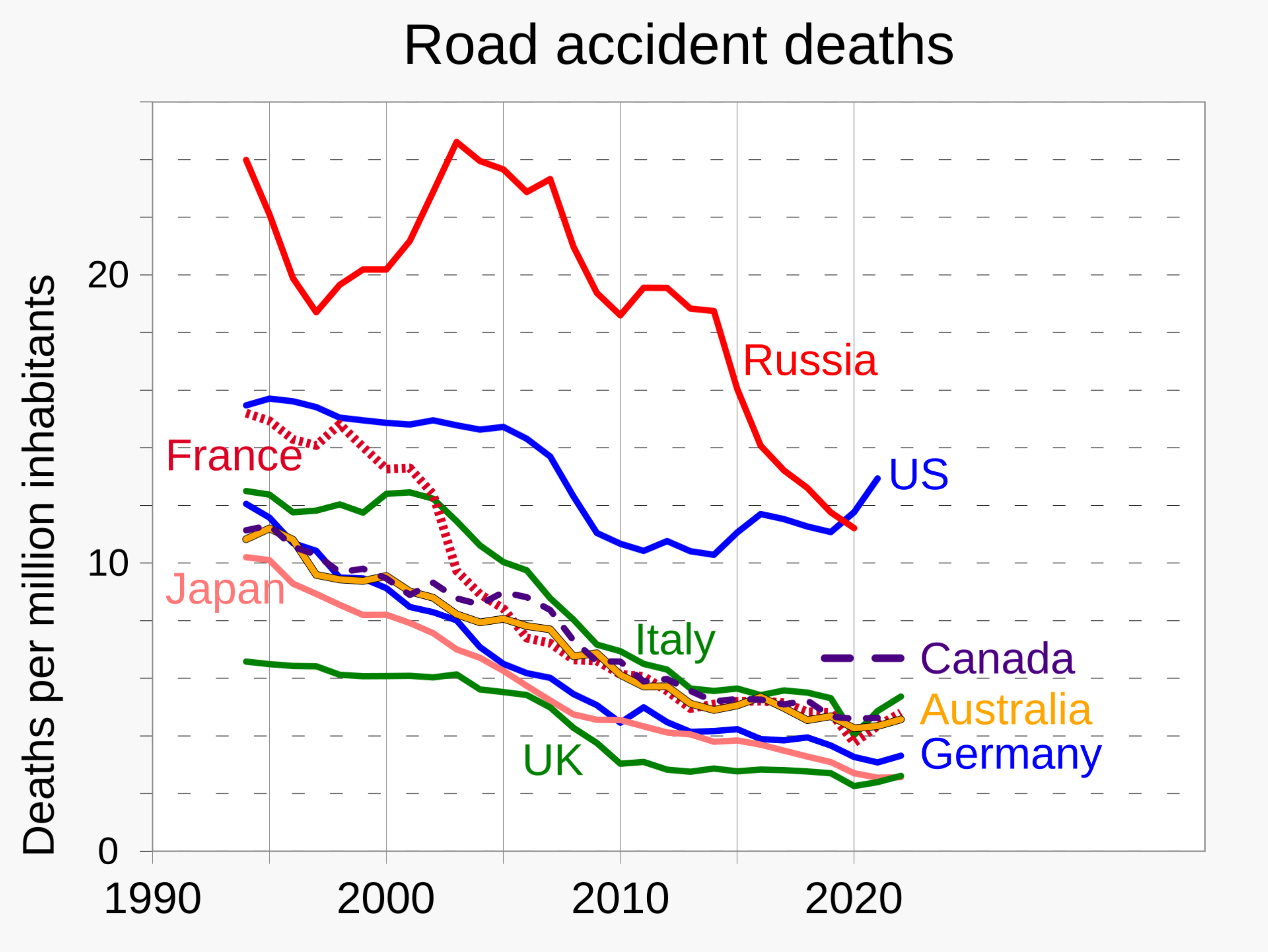
- Preliminaries
  1. Models
  2. Sensing
- Feature forecasting (EGO)
  1. Forecasting Depth and Flow
  2. Behavior Forecasting
- Memory based Trajectory forecasting (BEV)
  1. Forecasting w/o social context
  2. Socially-Aware forecasting
- Foundation models for Time Series
  1. Zero Shot Forecasting with LLMs
  2. Specialized Foundation Models



"Learn $F(x_{k-\delta:0};\theta)$ to model $p(x_{1:\tau}|\ x_{k-\delta:0})$"

Forecasting

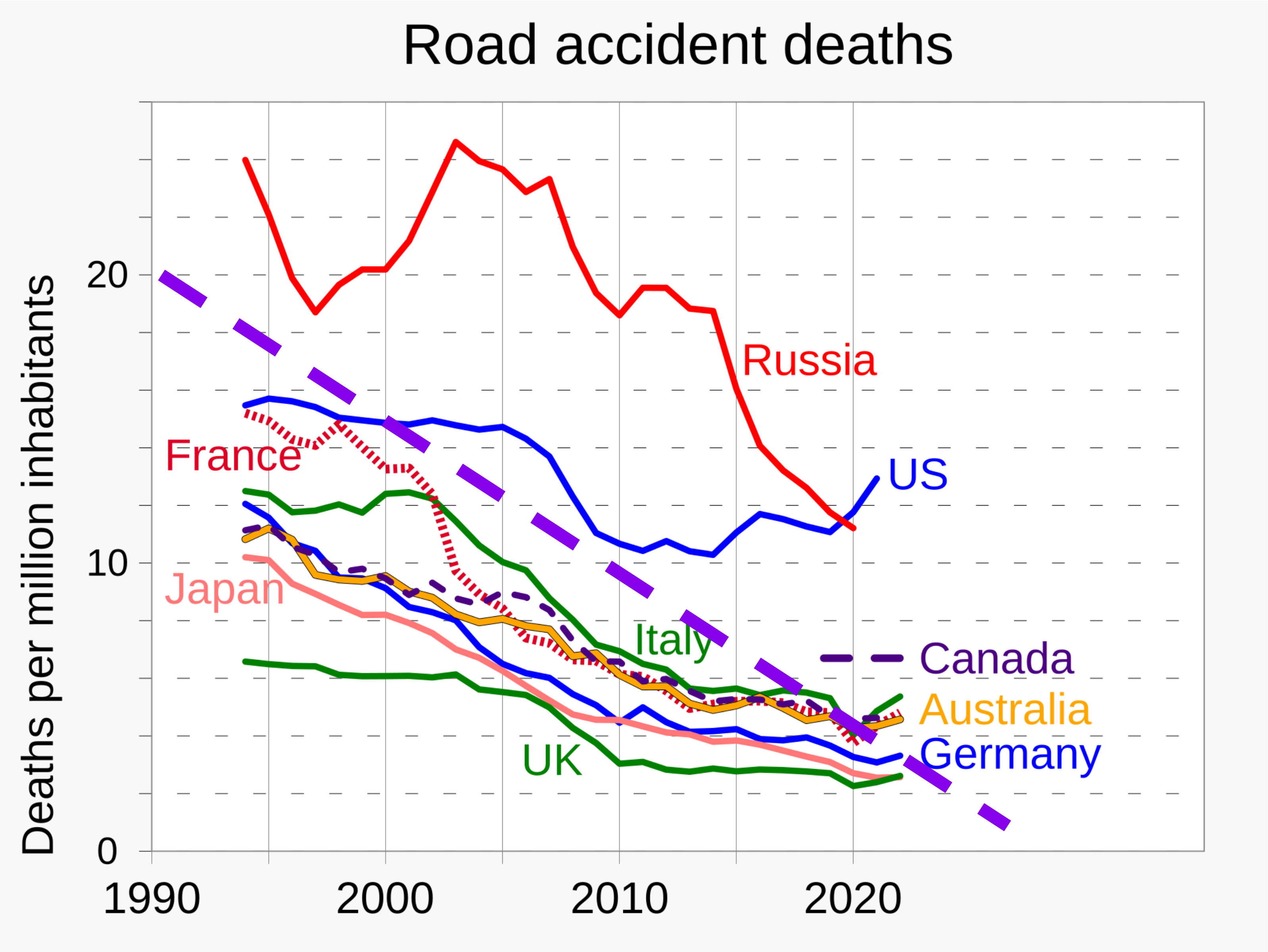
Motivation



- Per capita road accident deaths exhibit a slow decline over time
- Improved infrastructure, passive safety features, ADAS

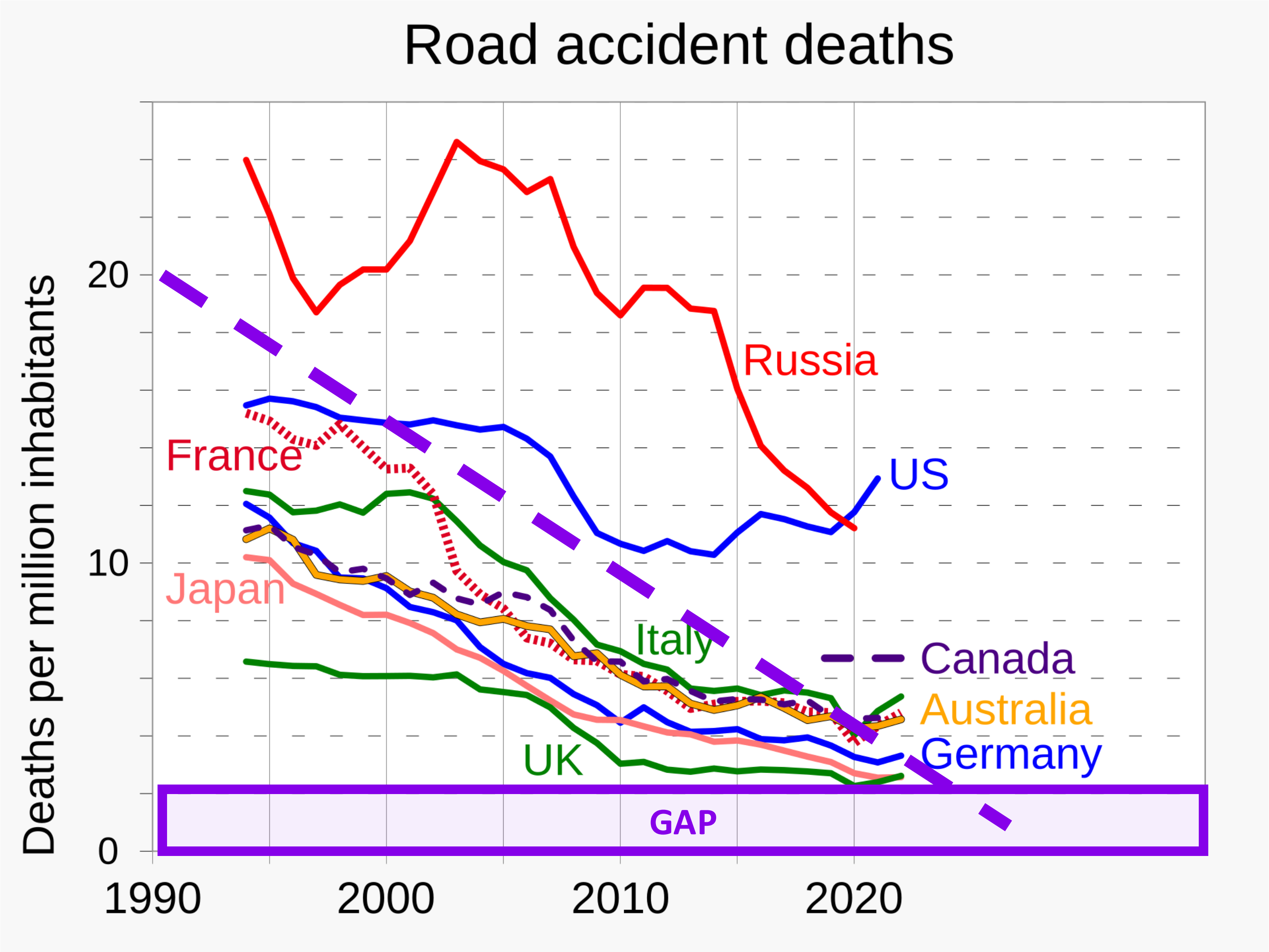
Motivation

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



- Per capita road accident deaths exhibit a slow decline over time
- Improved infrastructure, passive safety features, ADAS

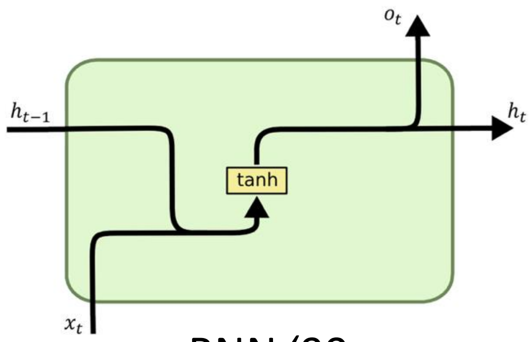
Motivation



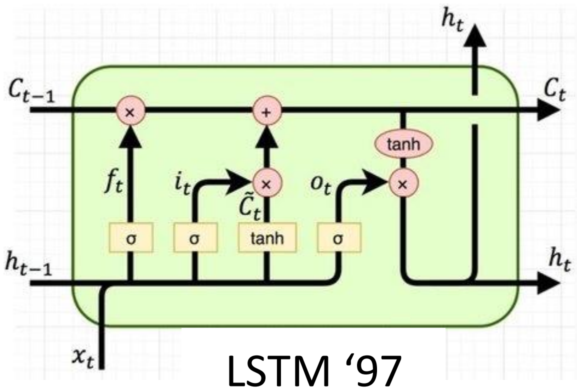
- Per capita road accident deaths exhibit a slow decline over time
- Improved infrastructure, passive safety features, ADAS
- Still not zero!

Preliminaries

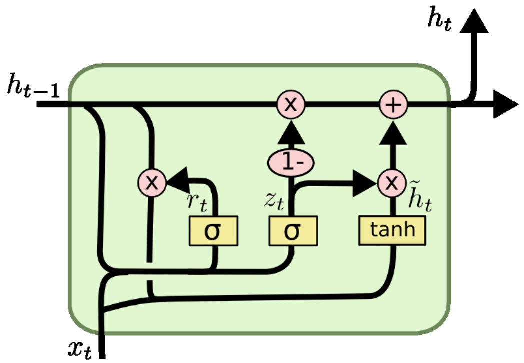
Models for Sequences*



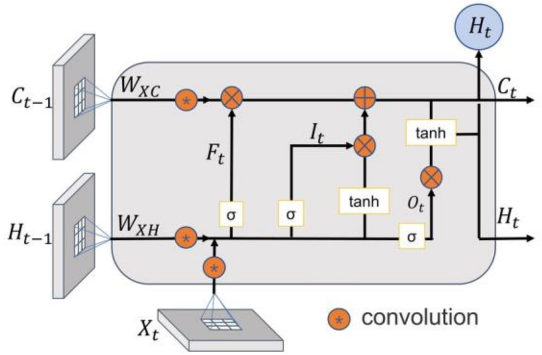
RNN ‘90



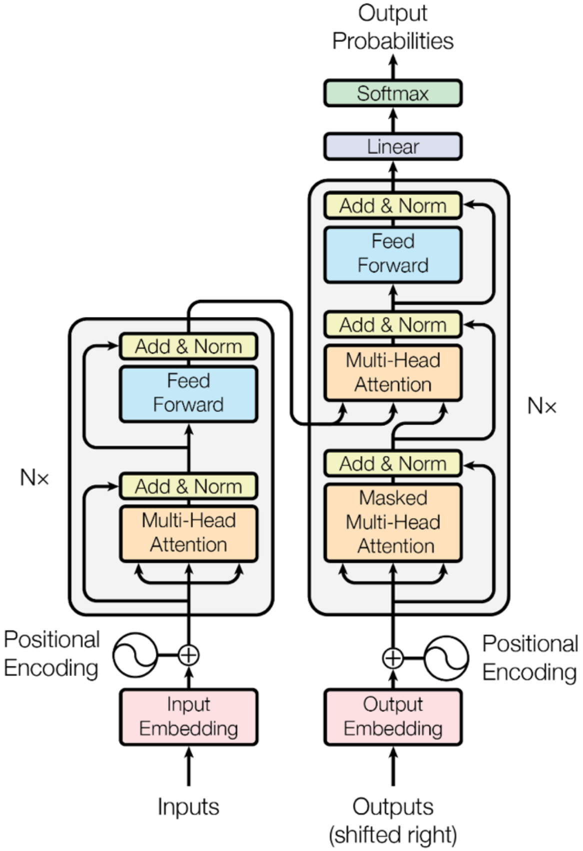
LSTM ‘97



GRU ‘14



ConvLSTM ‘15



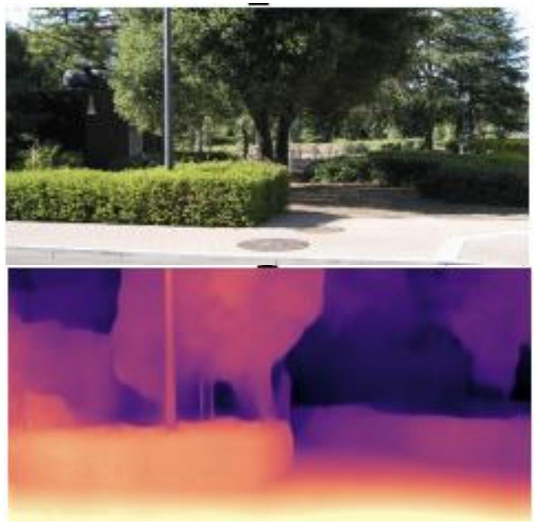
Transformers ‘17

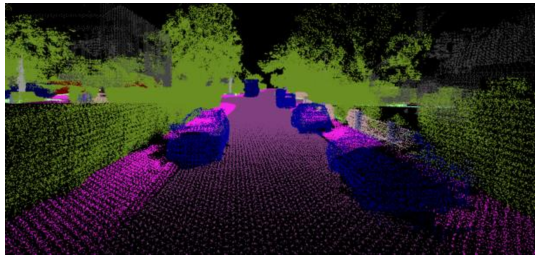
Elman, Jeffrey L. "Finding structure in time." Cognitive science 1990
Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 1997
Cho et al. “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation”. EMNLP 2014.
Shi, Xingjian, et al. "Convolutional LSTM network: A machine learning approach for precipitation nowcasting." NeurIPS 2015
Vaswani, Ashish, et al. "Attention is all you need." NeuIPS 2017

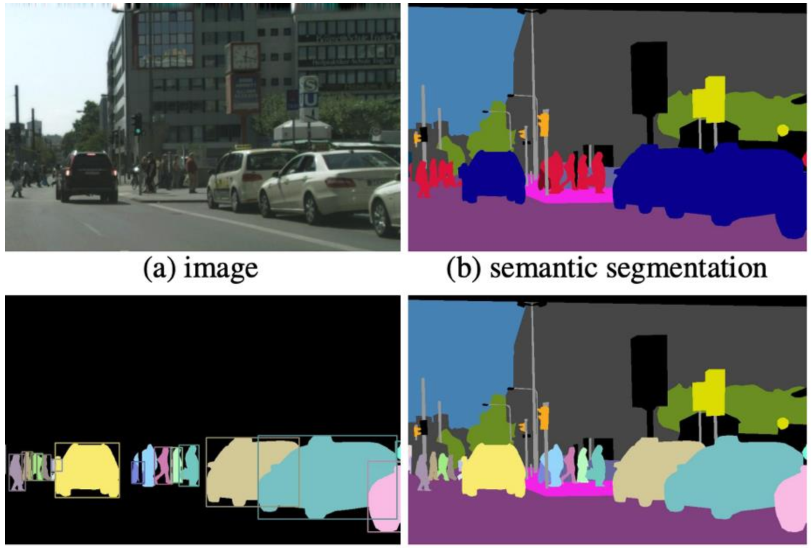
UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Sensing

- Standard computer vision tools can be exploited to gather information on:
  - moving objects in the scene (e.g.: Mask R-CNN, YOLOv11, DETR, SAM2)
  - Panoptic semantic scene understanding (e.g: Mask2Former)
  - 3D object location (e.g: LiDAR, stereo depth, DepthAnything, SLAM)





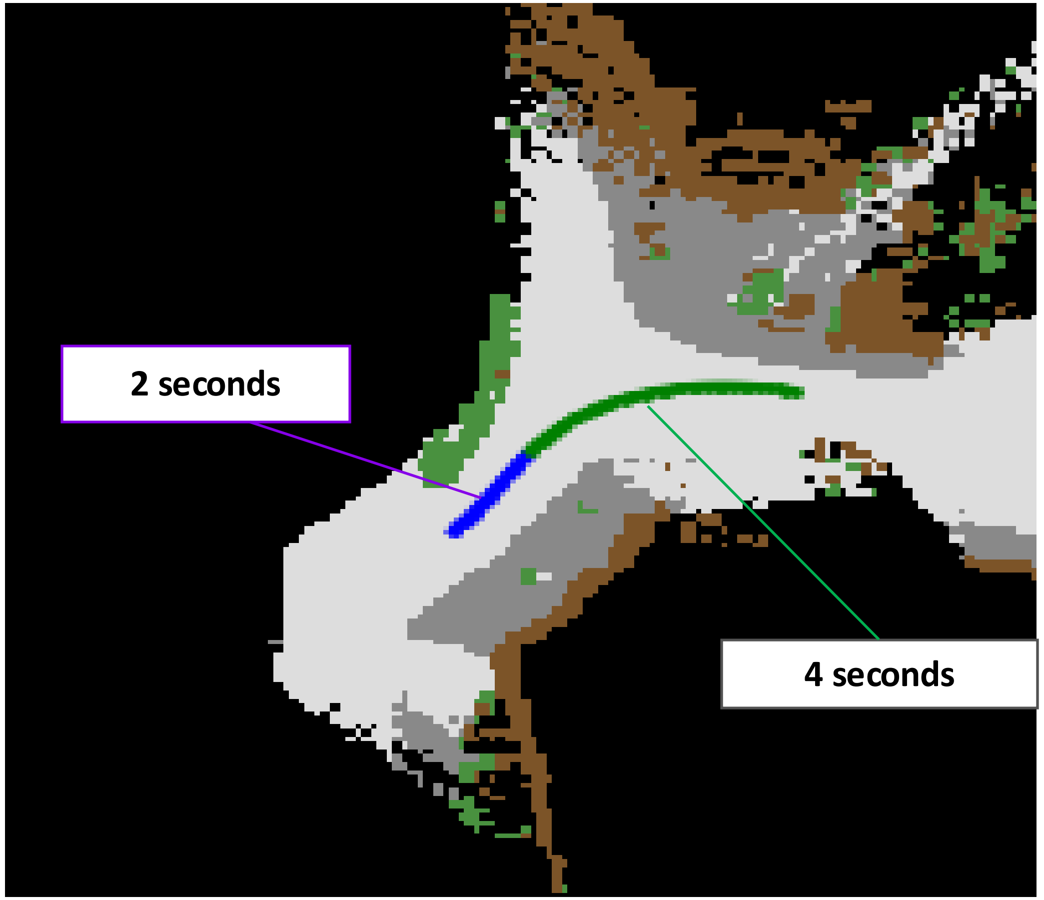


(a) image (b) semantic segmentation
(c) instance segmentation (d) panoptic segmentation

Ravi, N., et al. Sam 2: Segment anything in images and videos. ICLR 2025
Cheng, B., et al.. Masked-attention mask transformer for universal image segmentation. CVPR 2022
Yolov11 from Ultralitics - https://github.com/yt7589/yolov11
Yang, Lihe, et al. "Depth anything: Unleashing the power of large-scale unlabeled data." CVPR 2024.
Murai, et al. "MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors." CVPR 2025

Sensing

- Once sensing is performed, we work in a semantic top view map forecasting 2D trajectories so-called Bird's Eye View (BEV)

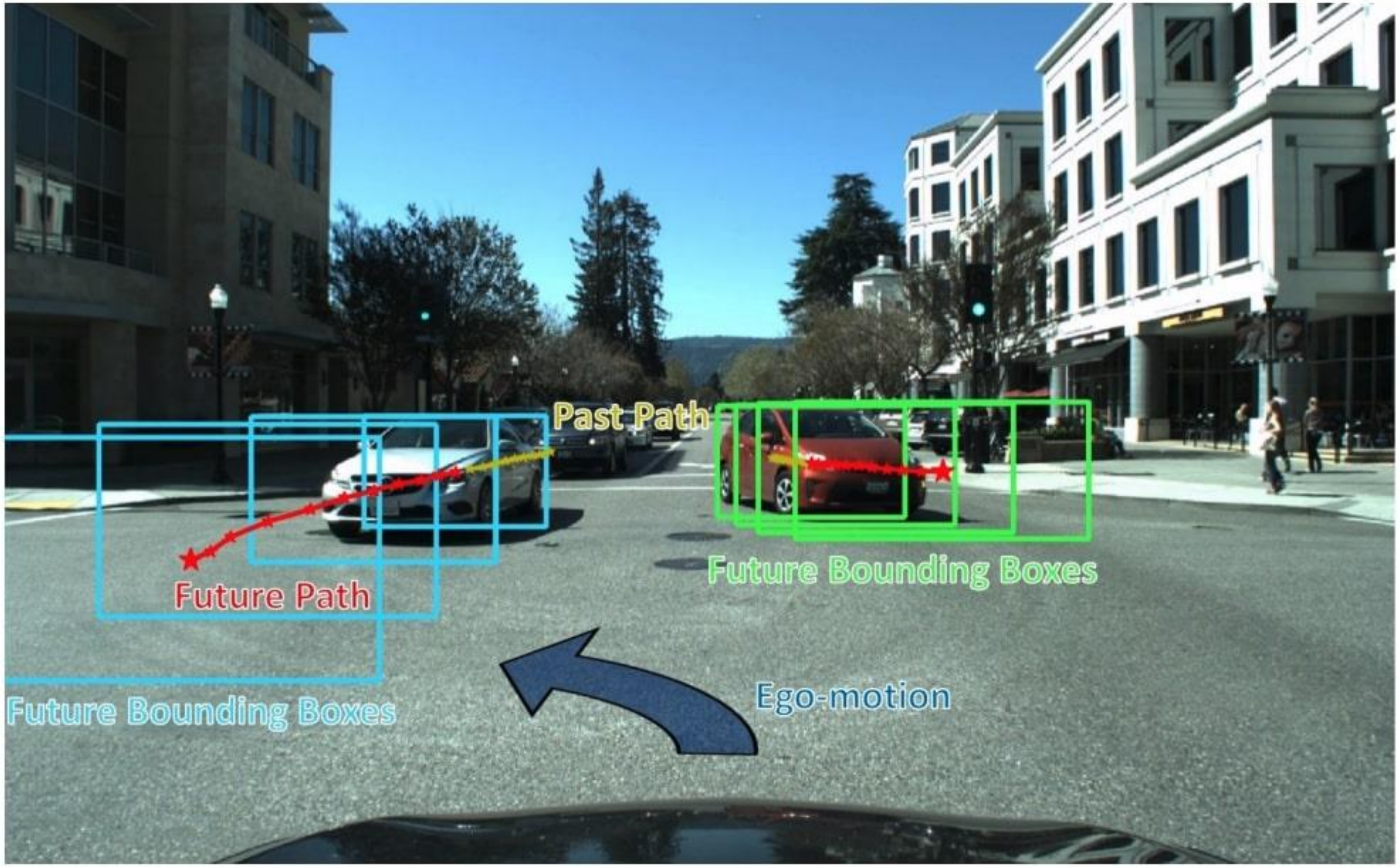


Ego Forecasting

Looking at the road

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Forecasting directly in image or feature space reduces the need for sensors and infrastructure
- Most general form of robotic perception pipeline
- Perform scene understanding in image space
- Two GOALs:
  - *Future object location*
  - *Future agent behavior*



Location Forecasting

Inferring segmentation

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Simple yet effective idea to forecast object location: predict the next frames autoregressively

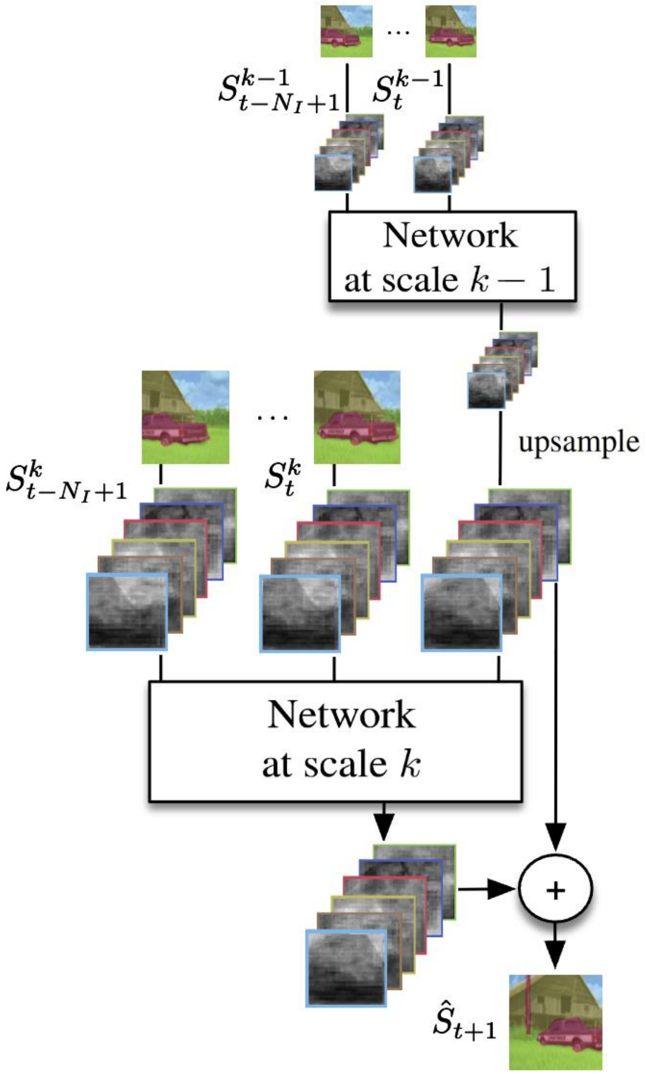
- Unfortunately forecasting RGB frames is extremely challenging (maybe Veo can help nowadays!)

- Solution: autoregressively predict *segmentation* from past segments!

$$\mathcal{L}(\hat{Y}, Y) = \mathcal{L}_{\ell_1}(\hat{Y}, Y) + \mathcal{L}_{\text{gdl}}(\hat{Y}, Y)$$

$$\sum_{i,j} \left| Y_{ij} - \hat{Y}_{ij} \right|$$

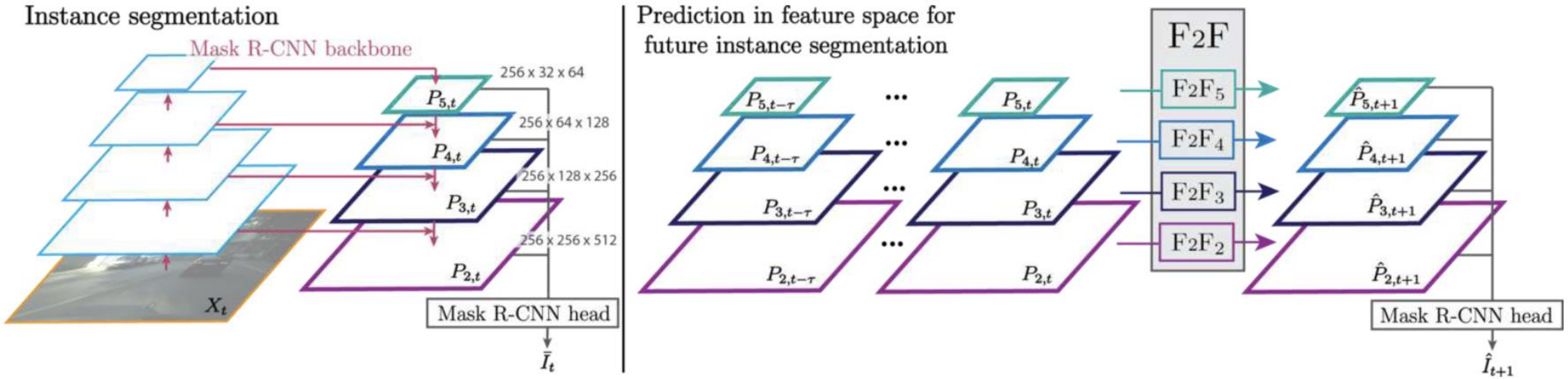
$$\sum_{i,j} \Big| \left| Y_{i,j} - Y_{i-1,j} \right| - \left| \hat{Y}_{i,j} - \hat{Y}_{i-1,j} \right| \Big| + \Big| \left| Y_{i,j-1} - Y_{i,j} \right| - \left| \hat{Y}_{i,j-1} - \hat{Y}_{i,j} \right| \Big|, \; ($$



Luc, Pauline, et al. "Predicting deeper into the future of semantic segmentation." ICCV 2017

Inferring features

- We can extract more information from the frames leveraging strong features instead of propagating segments
- Instead of predicting the segmentation, the model is trained to forecast intermediate representations
- A pre-trained detector is then applied as a *head* on such features providing future detections



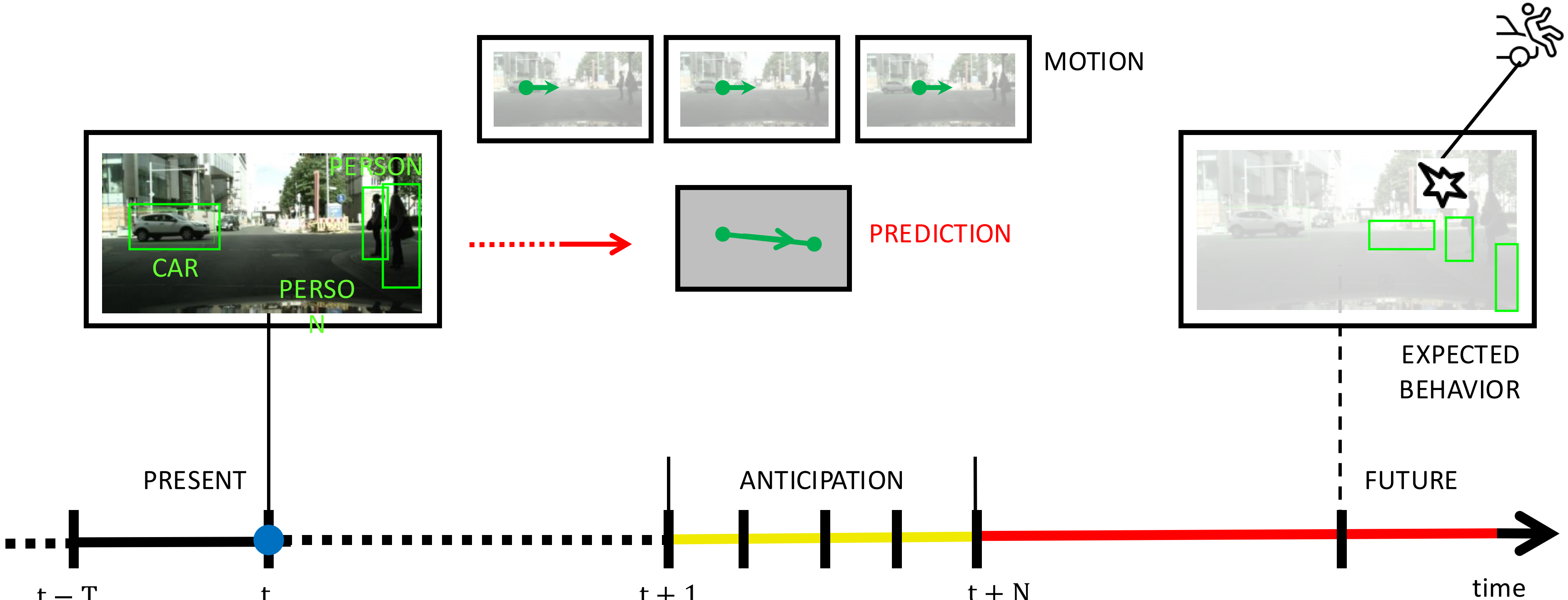
Luc, Pauline, et al. "Predicting future instance segmentation by forecasting convolutional features." ECCV 2018.

Inferring high level features

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

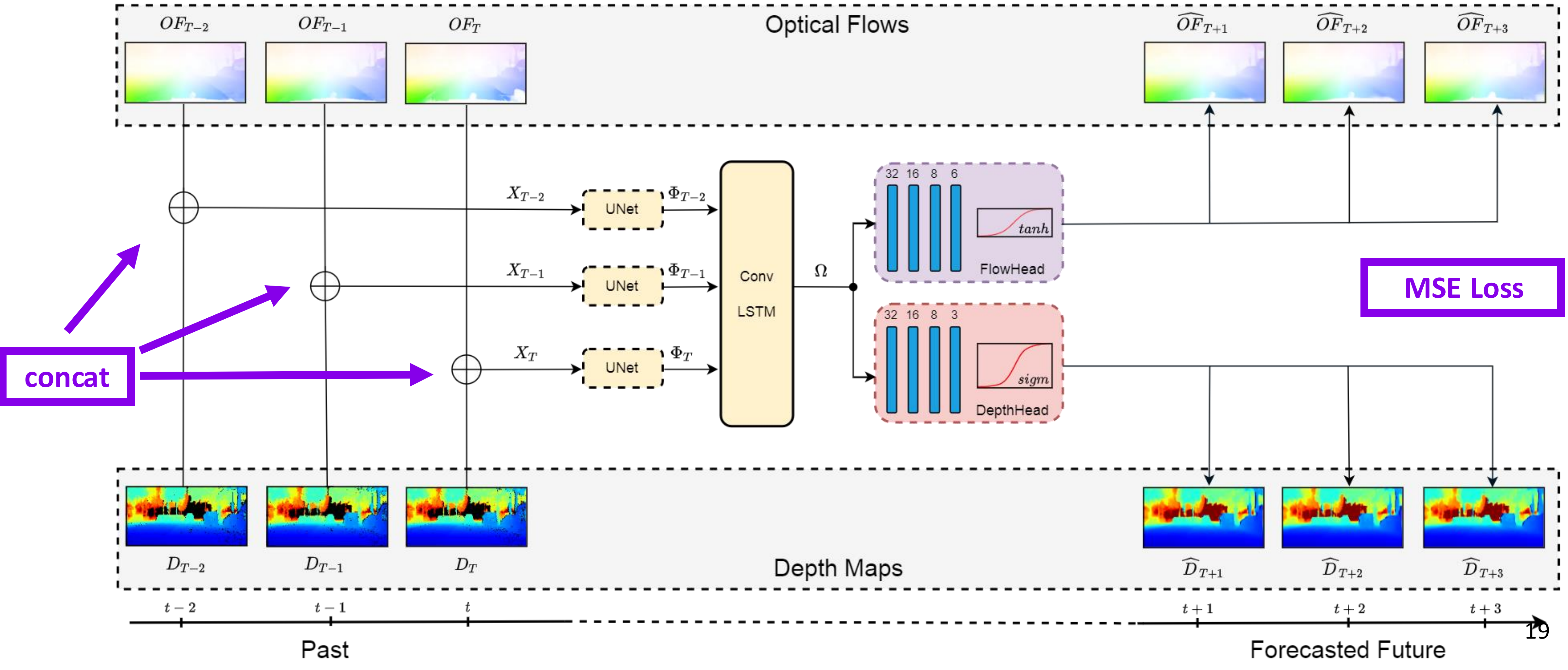
What give us a full understanding of a dynamic scene?

- *Optical flow* instantly tells where objects are headed
- *Depth* delivers 3D information
- Assuming object locations and classes known at time *t* can we forecast locations using Depth+Flow?



Inferring high level features

We design an architecture with the idea of feature sharing for the two tasks

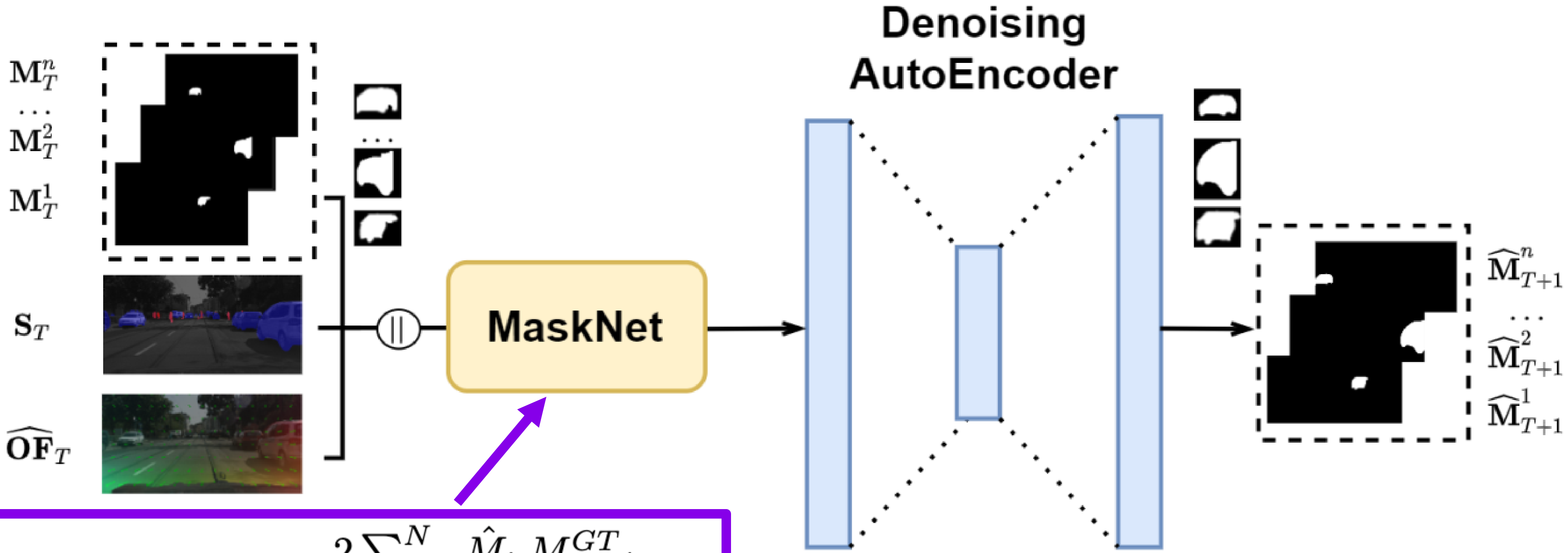


Ciamarra A., et al. FLODCAST: Flow and Depth Forecasting via Multimodal Recurrent Architectures. in Pattern Recognition. Elsevier. 2024

Learning to warp masks

- To predict future instances, we use MaskNet a Learned binary mask warper that learns to warp binary masks into the future given an initial segmentation and the cumulated flow

- A Denoising autoencoder is added downstream to improve the results



$$\mathcal{L}_{\text{mask}} = 1 - D = 1 - \frac{2\sum_{i=1}^{N} \hat{M}_i \, {M^{GT}}_i}{\sum_{i=1}^{N} \hat{M}_i^2 + \sum_{i=1}^{N} {M^{GT}}_i^2}$$

DICE Loss

Ciamarra A., et al. FLODCAST: Flow and Depth Forecasting via Multimodal Recurrent Architectures. in Pattern Recognition. Elsevier. 2024

Results

Can we infer future location of object more accurately with better future flow?

- Both Depth and Flow forecast get SOTA results (not reported here)
- MaskNet with simple flow forecasting gets SOTA on short term prediction
- We further improve thanks to the joint Depth+Flow prediction MaskNet results on Mid term

Method	Short term (T+3)			Mid term (T+9)		
	AP	AP50	IoU	AP	AP50	IoU
Mask RCNN oracle	34.6	57.4	73.8	34.6	57.4	73.8
MaskNet-Oracle [9]	24.8	47.2	69.6	16.5	35.2	61.4
Copy-last segm. [5]	10.1	24.1	45.7	1.8	6.6	29.1
Optical-flow shift [5]	16.0	37.0	56.7	2.9	9.7	36.7
Optical-flow warp [5]	16.5	36.8	58.8	4.1	11.1	41.4
Mask H2F [5]	11.8	25.5	46.2	5.1	14.2	30.5
F2F [5]	19.4	39.9	61.2	**7.7**	19.4	41.2
MaskNet [9]	**19.5**	**40.5**	**65.9**	6.4	18.4	45.5
MaskNet-FC	18.1	37.8	65.4	6.7	18.9	48.4
MaskNet-FC+DAE (Ours)	18.3	39.0	65.7	7.1	**20.7**	**49.2**

Ciamarra A., et al. FLODCAST: Flow and Depth Forecasting via Multimodal Recurrent Architectures. in Pattern Recognition. Elsevier. 2024

Funding & Collaborators

- Work done in collaboration with

Dr. Federico Becattini

Dr. Andrea Ciamarra

Prof. Alberto Del Bimbo

- Projects partially funded by

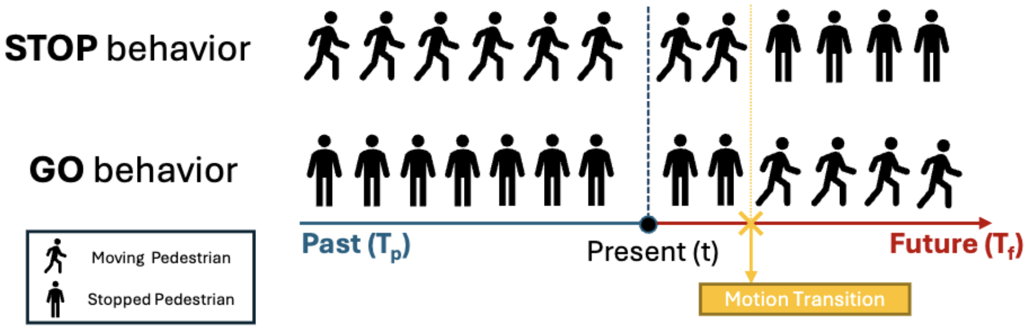
AI4media

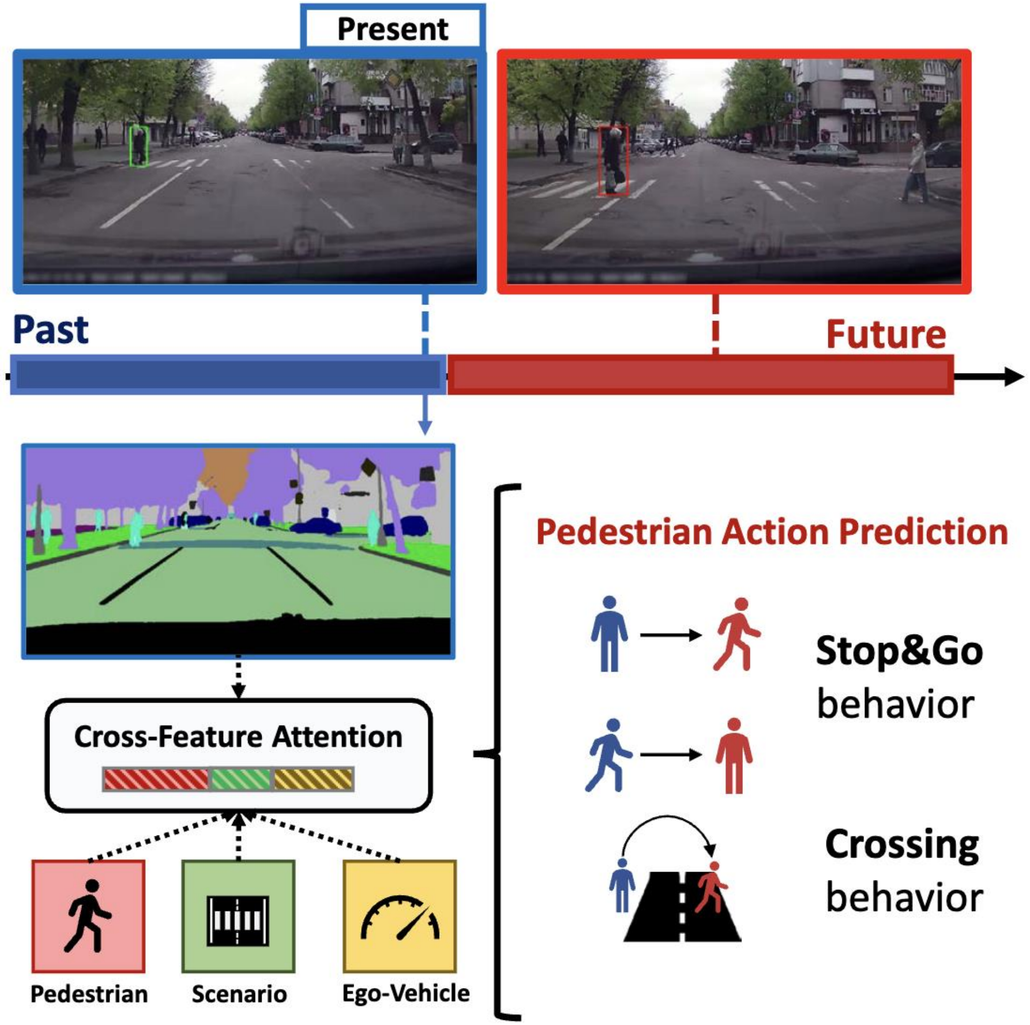
Ciamarra A., et al. FLODCAST: Flow and Depth Forecasting via Multimodal Recurrent Architectures. in Pattern Recognition. Elsevier. 2024

Behavior Forecasting

Behavior forecasting: what will they do?

- Usually focused on pedestrian behavior forecasting: stop, go or cross
- Stop/Go problems are defined as predicting if a person Stopped/Moving until time t will Go or Stop in a temporal window $[t, t + T_f]$
- All predictions can be made exploiting all observations available in the window $[t - T_p, t]$

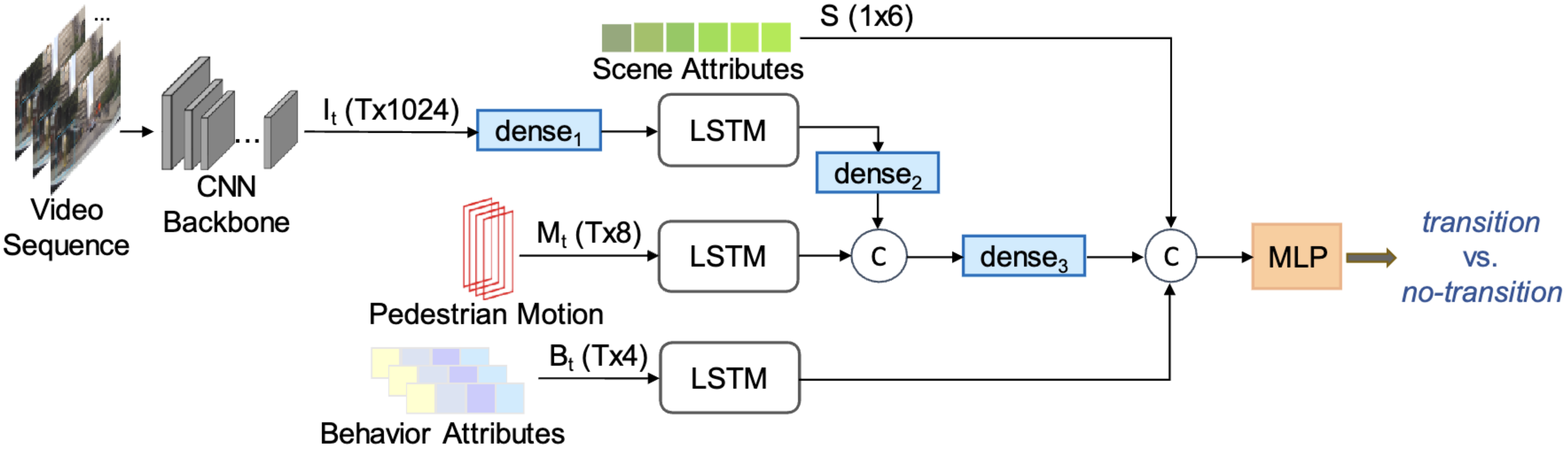




Marchetti, Francesco, et al. "CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting." IEEE Transactions on Intelligent Vehicles (2024).

Behavior forecasting: what will they do?

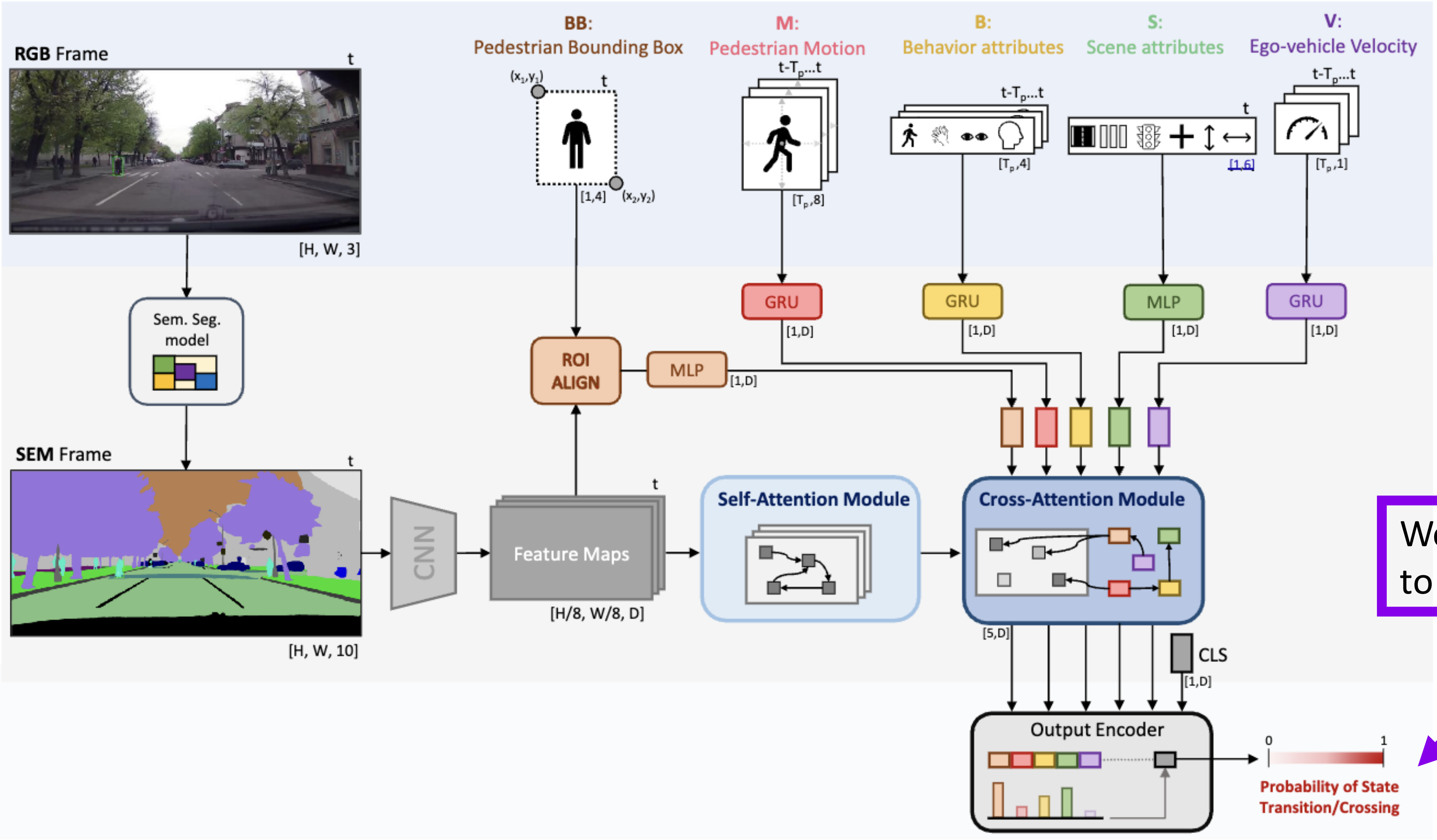
- First attempt: mix LSTM+CNN outputs for dynamic components and MLP for static components
- Late fusion with no self-attention or cross-attention



Guo, Dongxu, Taylor Mordan, and Alexandre Alahi. "Pedestrian stop and go forecasting with hybrid feature fusion." *ICRA 2022*

CrossFeat Architecture

- We leverage a transformer to efficiently blend diverse multimodal inputs



Weigthed BCE Loss to manage unbalance

Marchetti, Francesco, et al. "CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting." IEEE Transactions on Intelligent Vehicles (2024).

Results

- State-of-the art on the JAAD and TITAN behavior prediction

CrossFeat	Go			Stop		
	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Single query	66.6	66.8	62.3	55.6	60.1	60.5
Concatenation	83.6	66.8	68.3	69.8	67.8	65.7
Self-attention decoding	**88.9**	63.1	63.5	73.6	59.9	60.5
Complete	**88.9**	**68.1**	**70.1**	**75.4**	**71.0**	**67.3**

Benefit of fusion via transformer cross-attention

single frame

Model	Go			Stop		
	JAAD	PIE	TITAN	JAAD	PIE	TITAN
Static [18]	73.3	61.2	60.9	58.7	62.5	59.1
CrossFeat Static	**74.6**	60.4	**63.2**	**69.2**	**67.2**	**63.7**
Video [18]	76.4	64.7	62.9	62.9	64.2	61.7
Hybrid [18]	85.9	**70.2**	65.1	67.8	65.4	63.6
TED [21]	62.4	59.9	65.0	60.8	57.8	59.1
MTL [38]	62.0	63.3	64.5	67.6	59.6	56.7
CrossFeat (Ours)	**88.9**	68.1	**70.1**	**75.4**	**71.0**	**67.3**

multi frame

Stop/Go prediction State-of-the-art

Marchetti, Francesco, et al. "CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting." IEEE Transactions on Intelligent Vehicles (2024).

Funding & Collaborators

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Work done in collaboration with

Dr. Francesco Marchetti

Dr. Taylor Mordan

Dr. Federico Becattini

Prof. Alexandre Alahi

Prof. Alberto Del Bimbo

- Projects partially funded by

AI4media

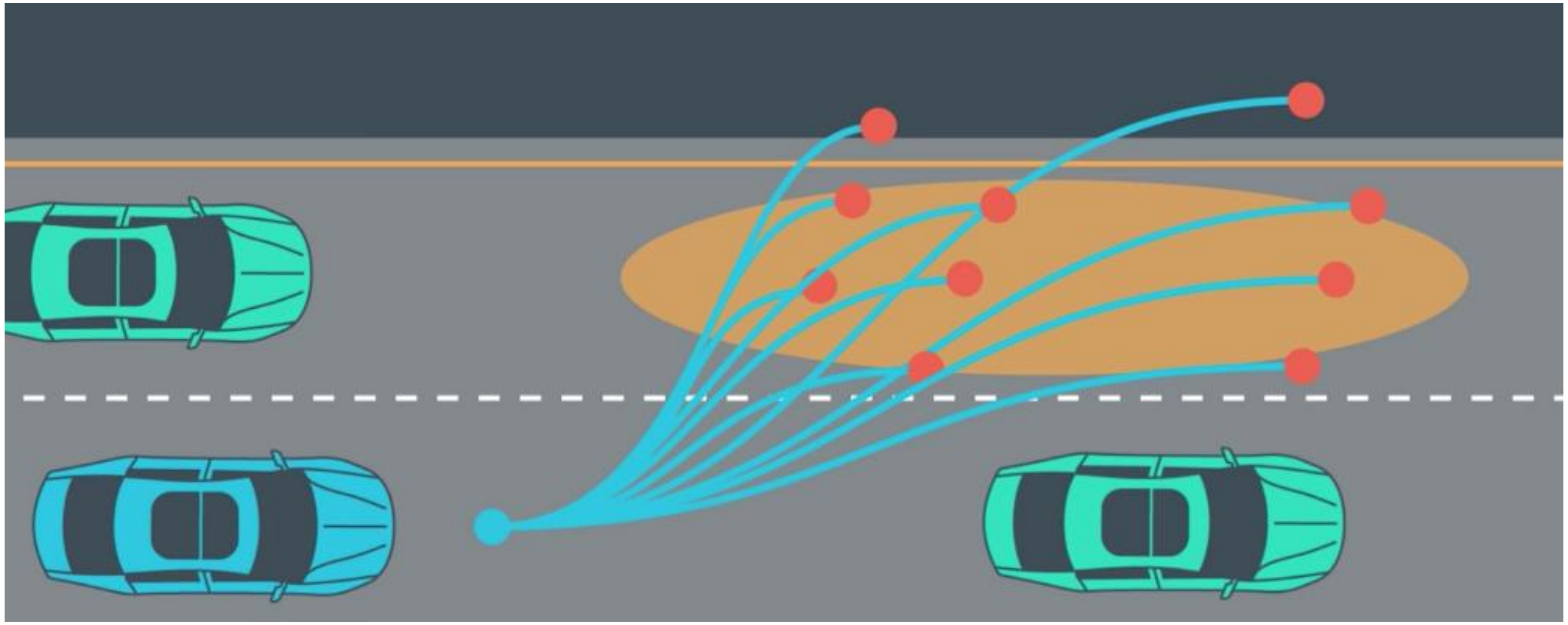
EPFL

Marchetti, Francesco, et al. "CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting." IEEE Transactions on Intelligent Vehicles (2024).

Trajectory Forecasting

Motivation

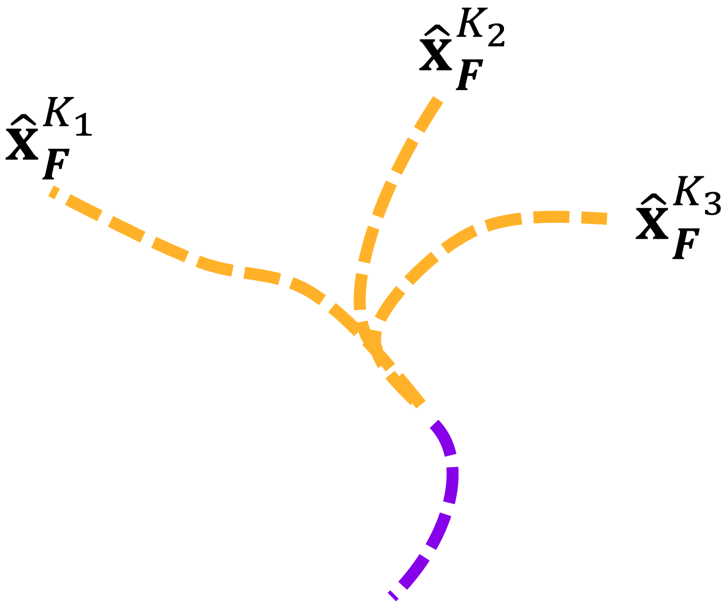
- Natural application to automotive: planning, collision avoidance, etc..



Trajectory Prediction

Problem definition

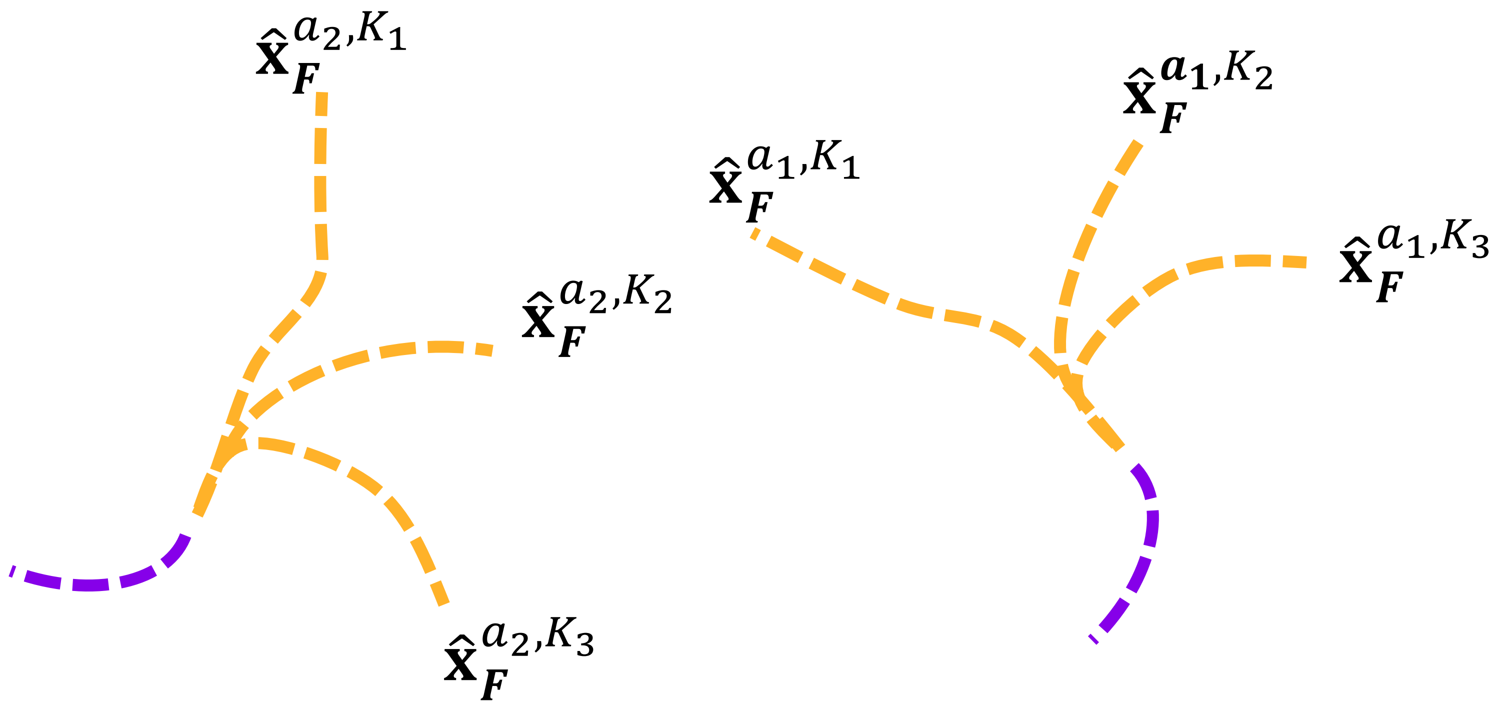
Given a set of previously observed locations $\boldsymbol{x}_{t-\tau}, \dots \boldsymbol{x}_t$ *in some state space (e.g.,* $\mathbb{R}^2$*), and some contextual information* $\boldsymbol{c}$ *predict* N $(K_1, \dots K_N)$ *multiple hypotheses of future locations* $\boldsymbol{x}^{i}_{t+1}, \dots, \boldsymbol{x}^{i}_{t+1+\Delta}$



Social Trajectory Prediction

Problem definition

Given a set of previously observed locations $\boldsymbol{x}_{t-\tau}, \ldots \boldsymbol{x}_t$ *in some state space (e.g.,* $\mathbb{R}^2$*), for a set of agents* A*, and some contextual information* $\boldsymbol{c}$*,* **jointly predict** N *multiple hypotheses of future locations* $\boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{t+1}}, \ldots, \boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{t+1+\Delta}}$ ***for each agent a***



Example

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Multiple futures are possible

State of the art

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- *Usage of a C-VAE to sample trajectories from a future distribution (e.g. DESIRE)*
- *Trajectories directly encoded in map representation combined with fully convolutional architectures (e.g. INFER)*
- *Social pooling modules to model interactions between different agents (e.g. Social-LSTM, Social-GAN)*
- *Goal based approaches to estimate trajectory endpoints (e.g. PECNet)*

- *Issues getting True multimodality + hard to manage the long tail*

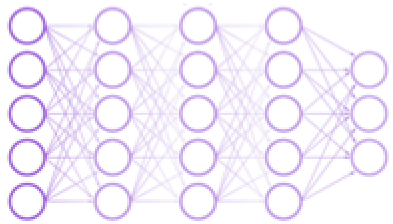
N. Lee et al. "Desire: Distant future prediction in dynamic scenes with interacting agents". CVPR 2017
S. Srikanth et al. "Infer: Intermediate representations for future prediction". IROS 2019
A. Alahi et al. "Social lstm: Human trajectory prediction in crowded spaces". CVPR 2016
A. Gupta et al. "Social gan: Socially acceptable trajectories with generative adversarial networks". CVPR 2018
K. Mangalam et al. "It is not the journey but the destination: Endpoint conditioned trajectory prediction". ECCV 2020

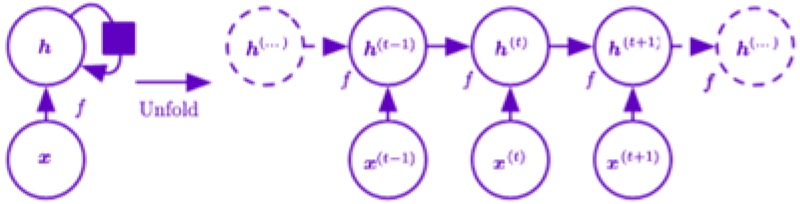
Memory Augmented Neural Networks

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Classical neural networks can be seen as *learnable functions*

- Depending on the domain/task we may design such architectures either with feed-forward structure or with a recurrent structure



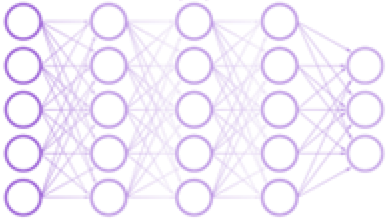


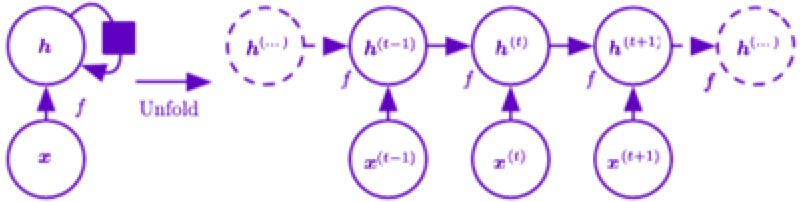
Memory Augmented Neural Networks

- Classical neural networks can be easily seen as *learnable functions*

- Here we rely on a **stateful** or **non-episodic** memory to augment the neural network

- We call our approach: **M**emory **A**ugmented **N**eural **TRA**jectory predictor: MANTRA





\+

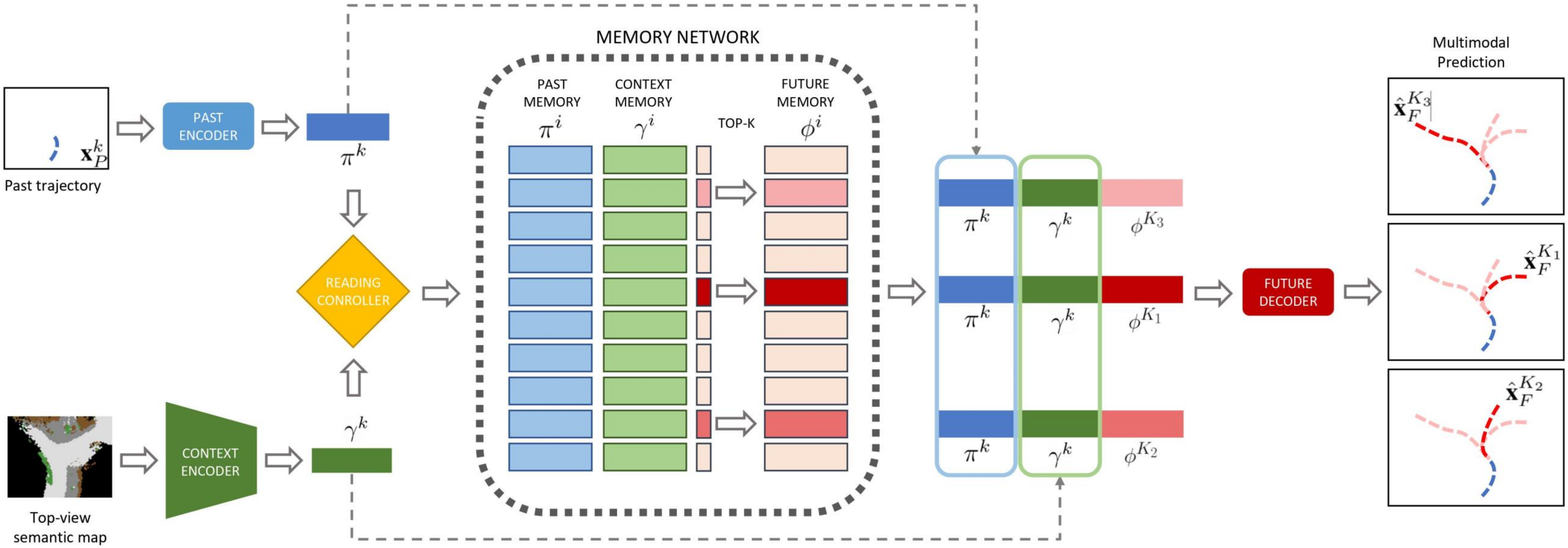
\+

Weston, Jason, Sumit Chopra, and Antoine Bordes. "Memory networks." *ICLR 2014*

MANTRA Overview

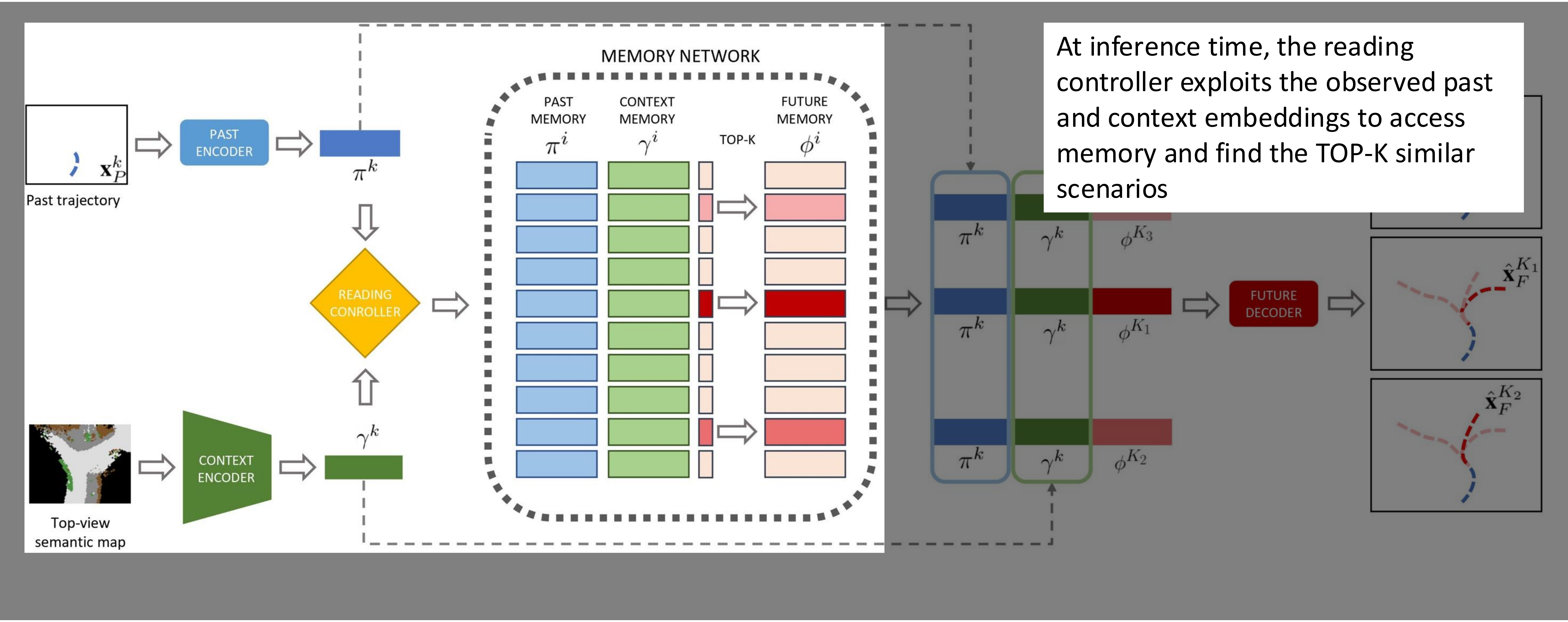
UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Memory is a feature store: must define how/when RW operations happen



F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo, *Mantra: Memory augmented networks for multiple trajectory prediction,* CVPR 2020

MANTRA Inference

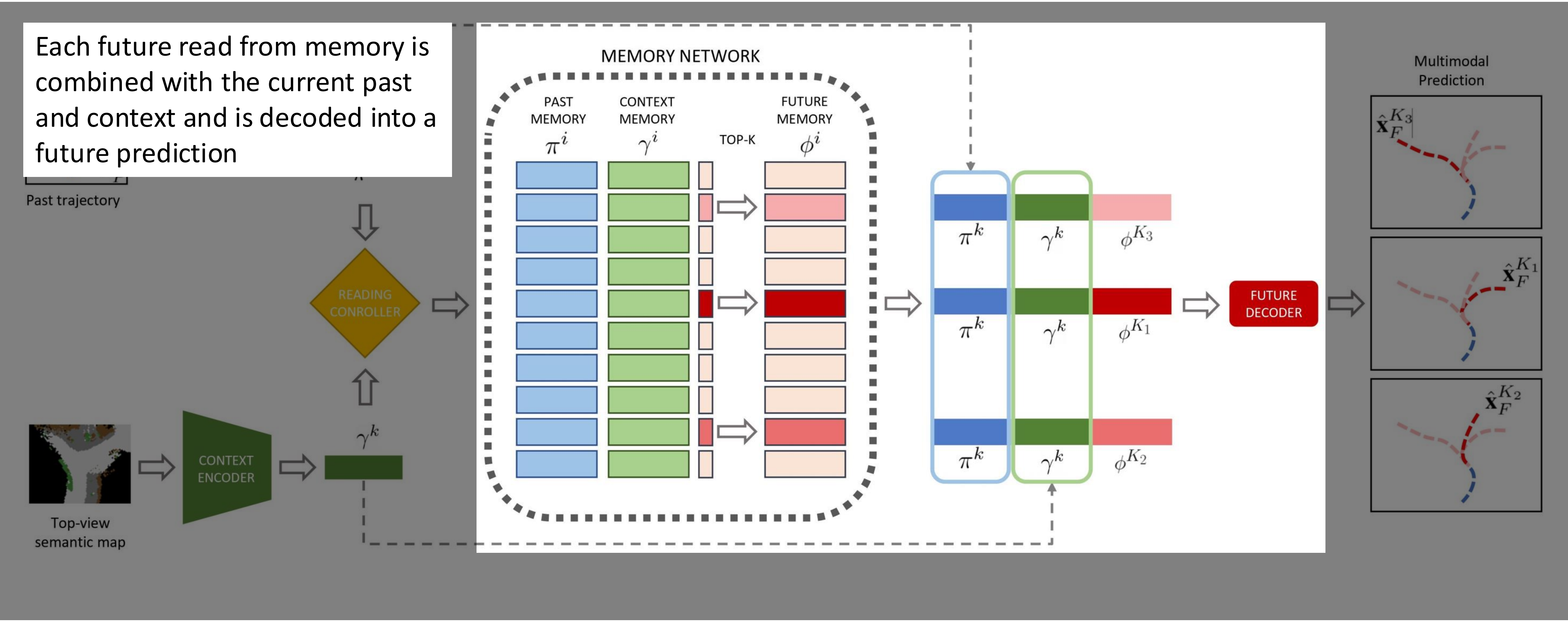


At inference time, the reading controller exploits the observed past and context embeddings to access memory and find the TOP-K similar scenarios

F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo, *Mantra: Memory augmented networks for multiple trajectory prediction,* CVPR 2020

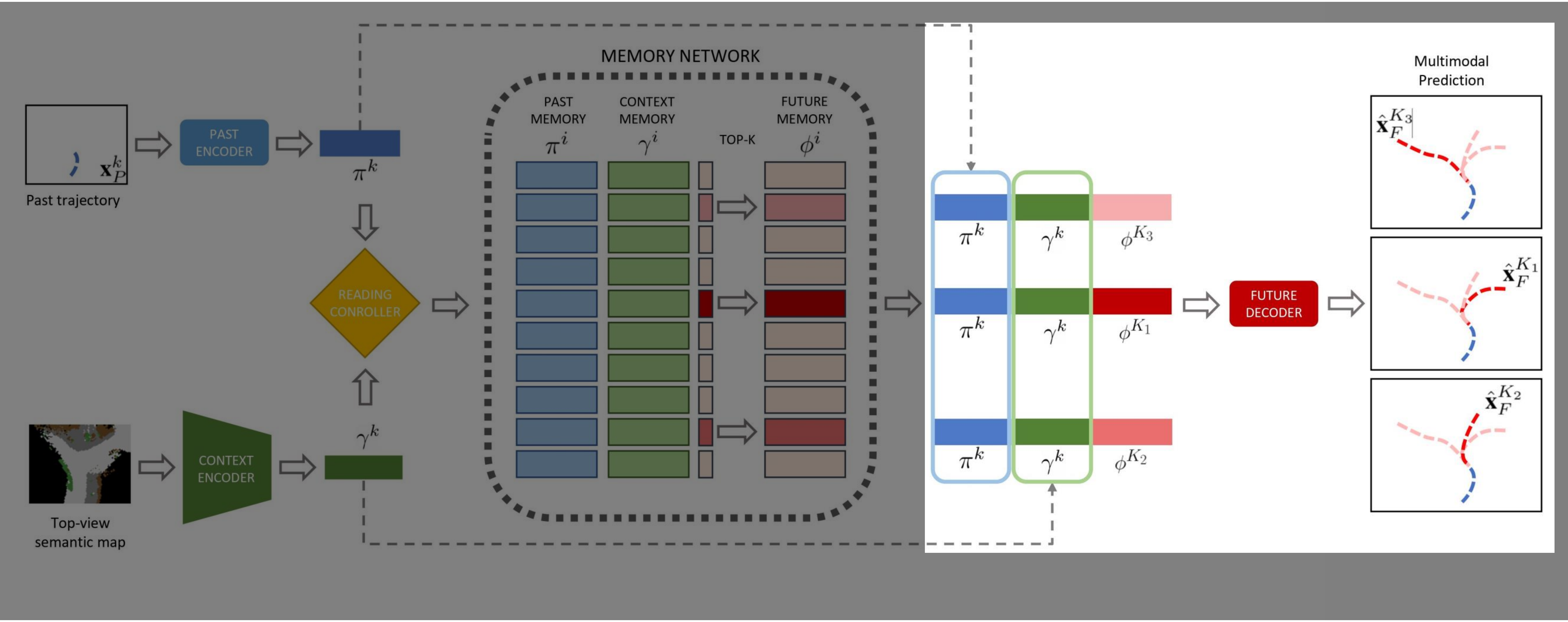
MANTRA Decoding

Each future read from memory is combined with the current past and context and is decoded into a future prediction



F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo, *Mantra: Memory augmented networks for multiple trajectory prediction,* CVPR 2020

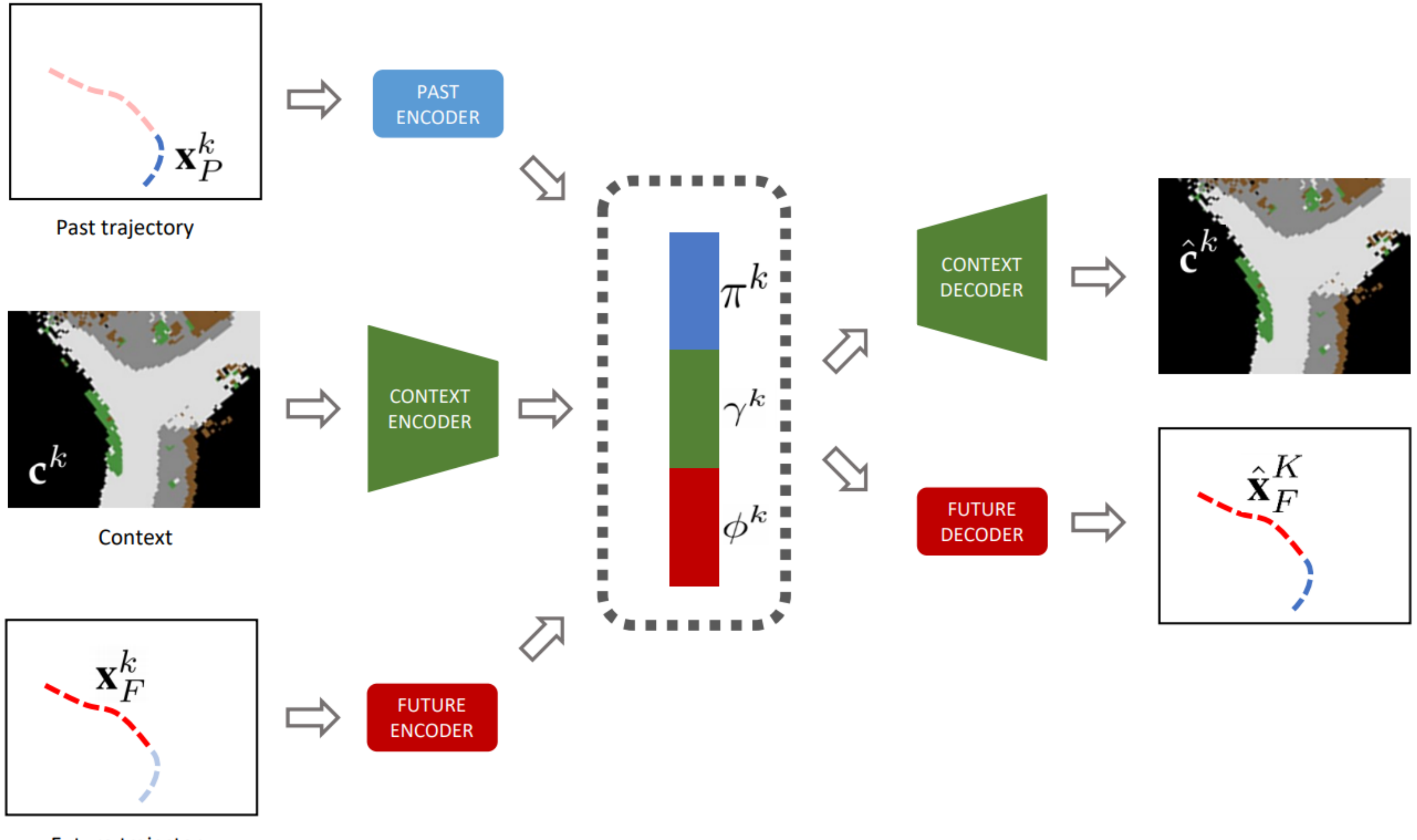
MANTRA Multimodal Prediction

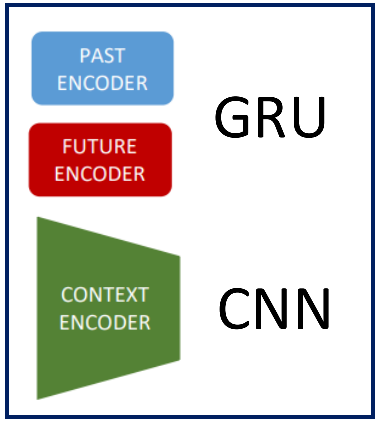


F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo, *Mantra: Memory augmented networks for multiple trajectory prediction,* CVPR 2020

Representation Learning

- To learn effective feature representations, we train encoding and decoding functions similarly to an autoencoder

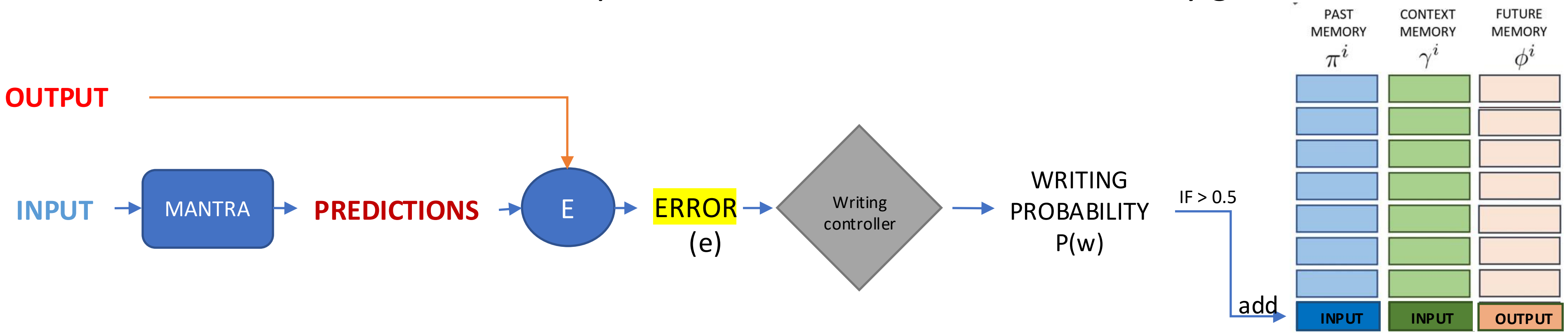




F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo, *Mantra: Memory augmented networks for multiple trajectory prediction,* CVPR 2020

Learning the writing controller

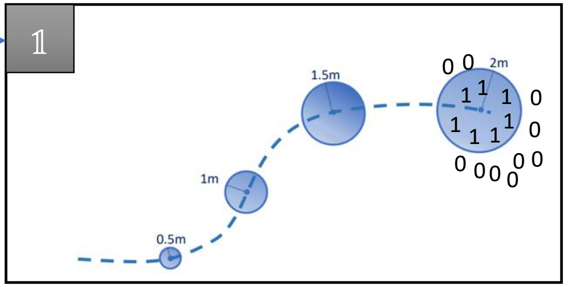
- The **writing** controller decides whether to insert a new example into memory **during training**
- Simple rule:
  - if prediction error is high with current memory: example should be stored
  - if the error is low, the example is not stored: the model is already good



Adaptive miss rate error

E

$$e = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_i(\hat{\mathbf{x}}_F, \mathbf{x}_F)$$



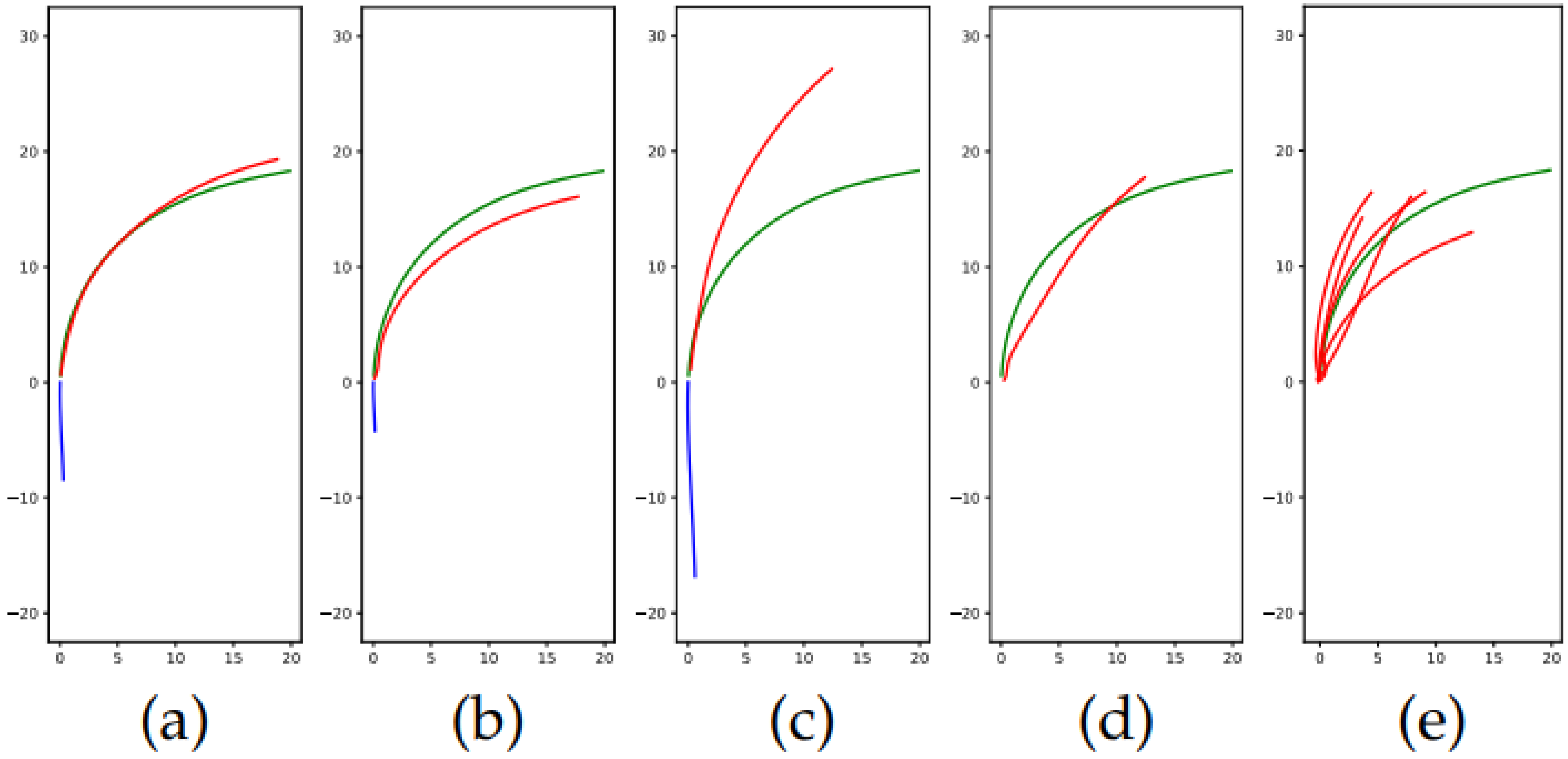
$$\mathcal{L}_w = e \cdot (1 - P(w)) + (1 - e) \cdot P(w)$$

Low error $\longrightarrow$ $\mathcal{L}_w \approx P(w)$

High error $\longrightarrow$ $\mathcal{L}_w \approx 1 - P(w)$

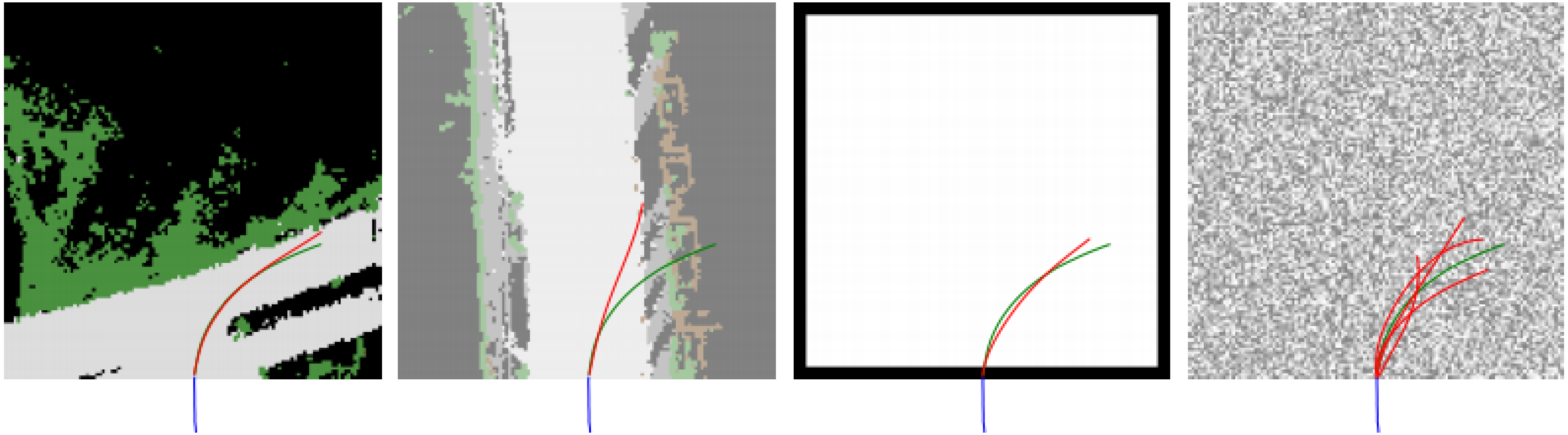
Decoder Analysis: past

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



observed past | slower past | faster past | zeroed past | randomized pasts

Decoder Analysis: context



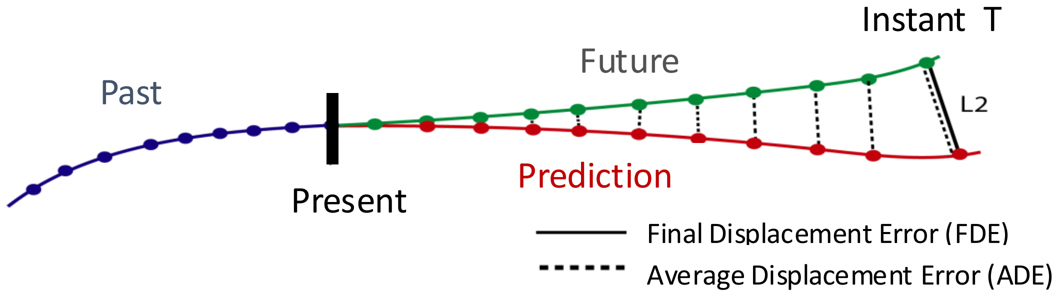
(a) (b) (c) (d)

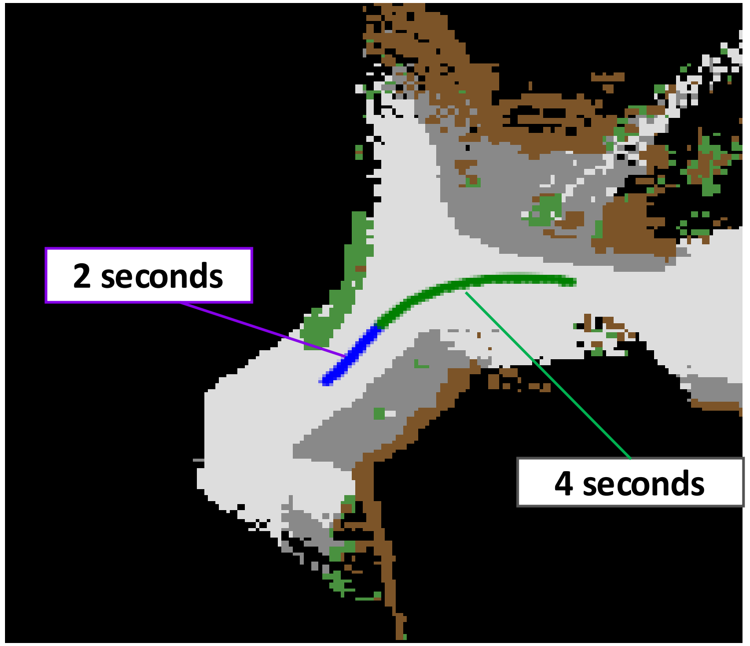
CONTEXT: Original | different | Embedding zeroed | Multiple randomized embeddings

Dataset and metrics

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Two datasets: KITTI (10k tracks) and ARGOVERSE (300k tracks)
- Metrics: Average Displacement Error (ADE)
  Final Displacement Error (FDE)





KITTI

Method	ADE 1s	ADE 2s	ADE 3s	ADE 4s	FDE 1s	FDE 2s	FDE 3s	FDE 4s
Kalman	0.33	0.54	0.93	1.4	0.46	1.18	2.18	3.32
Linear	0.31	0.56	0.89	1.28	0.47	1.13	1.94	2.87
MLP	0.30	0.54	0.88	1.28	0.46	1.12	1.94	2.88
RNN Enc-Dec [78]	0.68	1.94	3.20	4.46	-	-	-	-
Markov [52]	0.70	1.41	2.12	2.99	-	-	-	-
Conv-LSTM (top 5) [52]	0.76	1.23	1.60	1.96	-	-	-	-
INFER (top 1) [52]	0.75	0.95	1.13	1.42	1.01	1.26	1.76	2.67
INFER (top 5) [52]	0.56	0.75	0.93	1.22	0.81	1.08	1.55	2.46
MANTRA (top 1)	0.37	0.67	1.07	1.55	0.60	1.33	2.32	3.50
MANTRA (top 5)	0.33	0.48	0.66	0.90	0.45	0.78	1.22	2.03
MANTRA (top 10)	0.31	0.43	0.57	0.78	0.43	0.67	1.04	1.78
MANTRA (top 20)	**0.29**	**0.41**	**0.55**	**0.74**	**0.41**	**0.64**	**1.00**	**1.68**

Argoverse

Method	ADE 1s	ADE 3s	FDE 1s	FDE 3s
Kalman (top 1)	0.72	2.70	1.29	6.56
Linear (top 1)	0.58	1.95	0.98	4.58
MLP (top 1)	0.53	1.68	0.87	3.90
NN [1] (top 1)	0.75	2.46	1.28	5.60
NN + map [1] (top 6)	0.72	2.28	1.33	4.80
LSTM ED [1] (top 1)	0.68	2.27	1.78	5.19
LSTM ED + map [1] (top 6)	0.80	2.25	1.35	4.67
MFP [9] (top 6)	-	1.39	-	-
MANTRA (top 1)	0.72	2.36	1.25	5.31
MANTRA (top 6)	0.56	1.22	0.84	2.30
MANTRA (top 10)	0.53	1.00	0.77	1.69
MANTRA (top 20)	**0.52**	**0.84**	**0.73**	**1.16**

Won an Honorable Mention at the Argoverse Challenge hosted during the WAD workshop(CVPR2020)

Zero Shot transfer

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- MANTRA zero-shot transfer capability: training on KITTI and evaluation on Cityscapes and Oxford RobotCar

Oxford RobotCar

Method	ADE				FDE			
	1s	2s	3s	4s	1s	2s	3s	4s
INFER (top 1) [8]	1.06	1.35	1.48	1.68	1.31	1.71	1.70	2.56
INFER (top 5) [8]	0.85	1.14	1.29	1.50	1.18	1.58	1.58	2.41
MANTRA (top 1)	0.55	0.77	1.01	1.30	0.60	1.15	1.82	2.63
MANTRA (top 5)	0.55	0.68	0.82	1.03	0.58	0.88	1.37	2.07
MANTRA (top 10)	0.44	0.56	0.72	0.94	0.48	0.73	1.33	1.98
MANTRA (top 20)	**0.31**	**0.43**	**0.59**	**0.83**	**0.35**	**0.61**	**1.24**	**1.96**

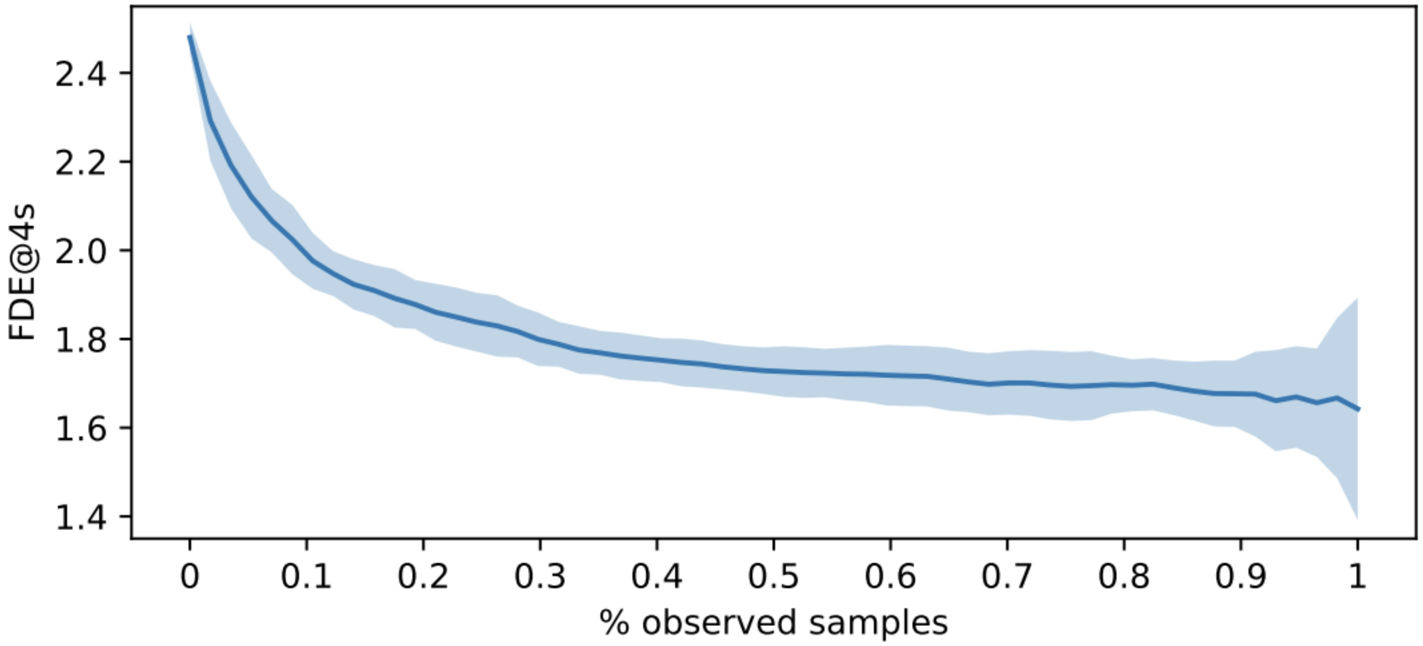
Cityscapes

Method	ADE	FDE
Conv-LSTM (top 1) [8]	1.50	-
Conv-LSTM (top 3) [8]	1.36	-
Conv-LSTM (top 5) [8]	1.28	-
INFER (top 1) [8]	1.11	1.59
INFER (top 3) [8]	0.99	1.45
INFER (top 5) [8]	0.91	1.38
MANTRA (top 1)	0.81	1.42
MANTRA (top 3)	0.66	1.15
MANTRA (top 5)	0.60	1.00
MANTRA (top 10)	0.54	0.86
MANTRA (top 20)	**0.49**	**0.79**

Incremental Setting

UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

- We test MANTRA incremental learning capabilities
- The model observes batches of test samples online, that are used as training data
- MANTRA is evaluated on the remaining portion of the test set.



Code Available!

UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

https://github.com/Marchetz/MANTRA-CVPR20

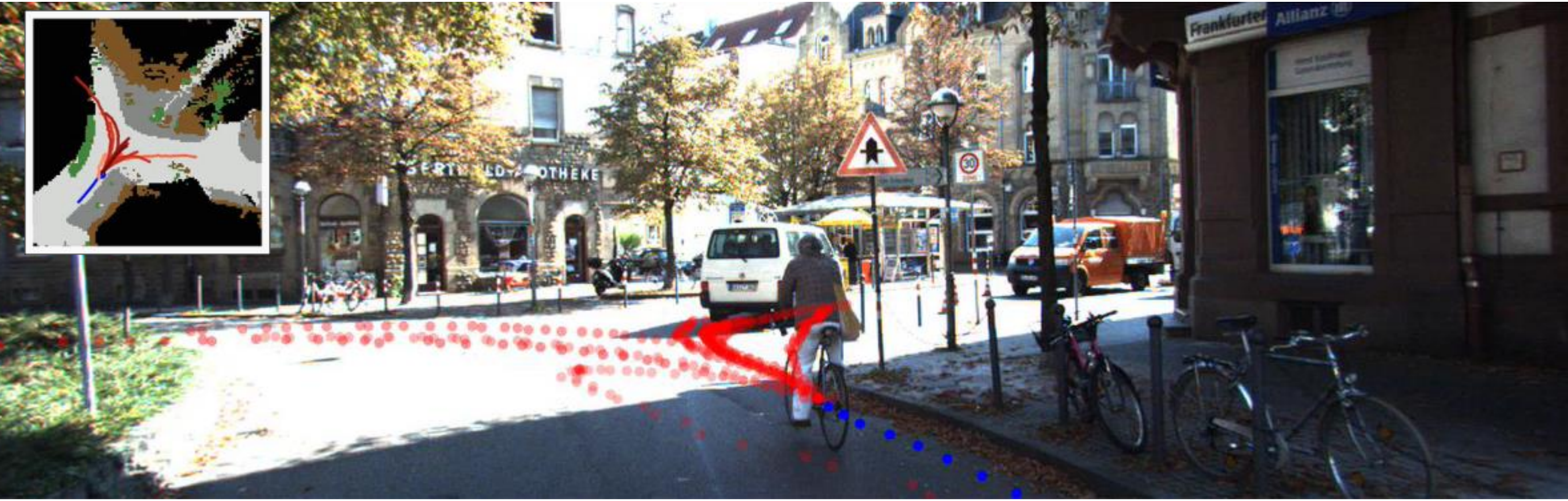
References

F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo, *Mantra: Memory augmented networks for multiple trajectory prediction,* CVPR 2020

F. Marchetti, F. Becattini, L. Seidenari, A. Del Bimbo, *Multiple Trajectory Prediction of Moving Agents with Memory Augmented Networks,* PAMI 2021

Federico Becattini, Francesco Marchetti, Lorenzo Seidenari, Alberto Del Bimbo, ABAD Frédéric, Kévin Buchicchio, Rémy Bendahan, Publication date, 2023/4/27, App. No. 17928163

Trajectory Forecasting Example

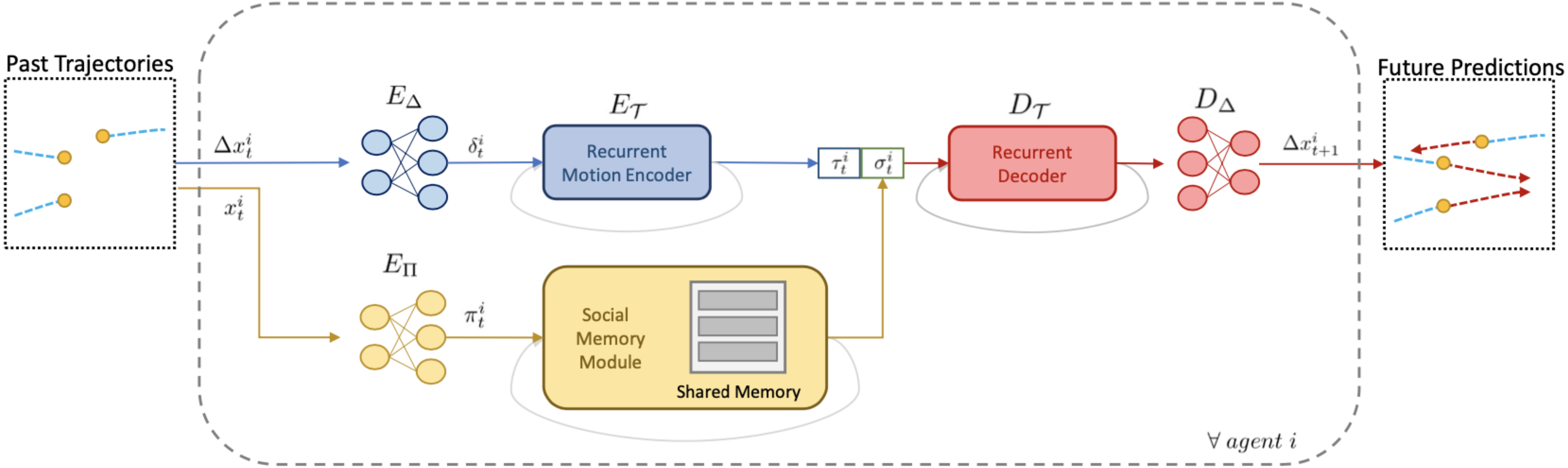


Socially-Aware Forecasting

Social Memory

- MANTRA lacks a social model.

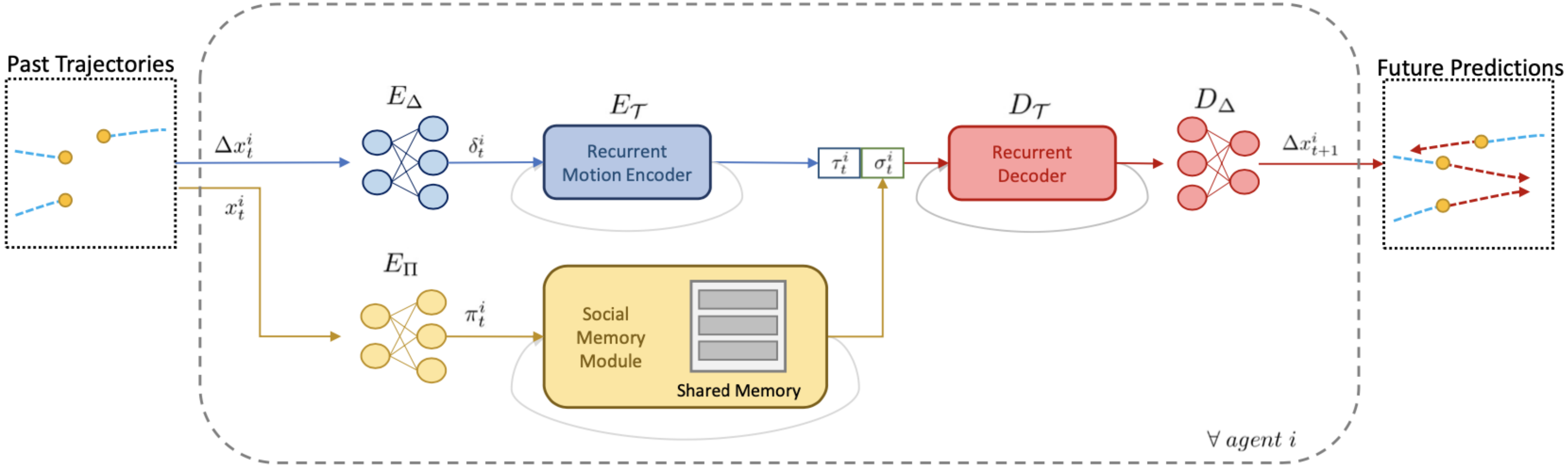
- SMEMO employs a shared memory to allow awareness in agent future trajectory prediction



Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

Social Memory

- SMEMO is an **end-to-end** trainable **episodic** memory augmented neural network
- At each time step to predict any agent SMEMO can leverage information **stored into the memory**
- Each agent is responsible for updating (read/write) the memory during an episode.

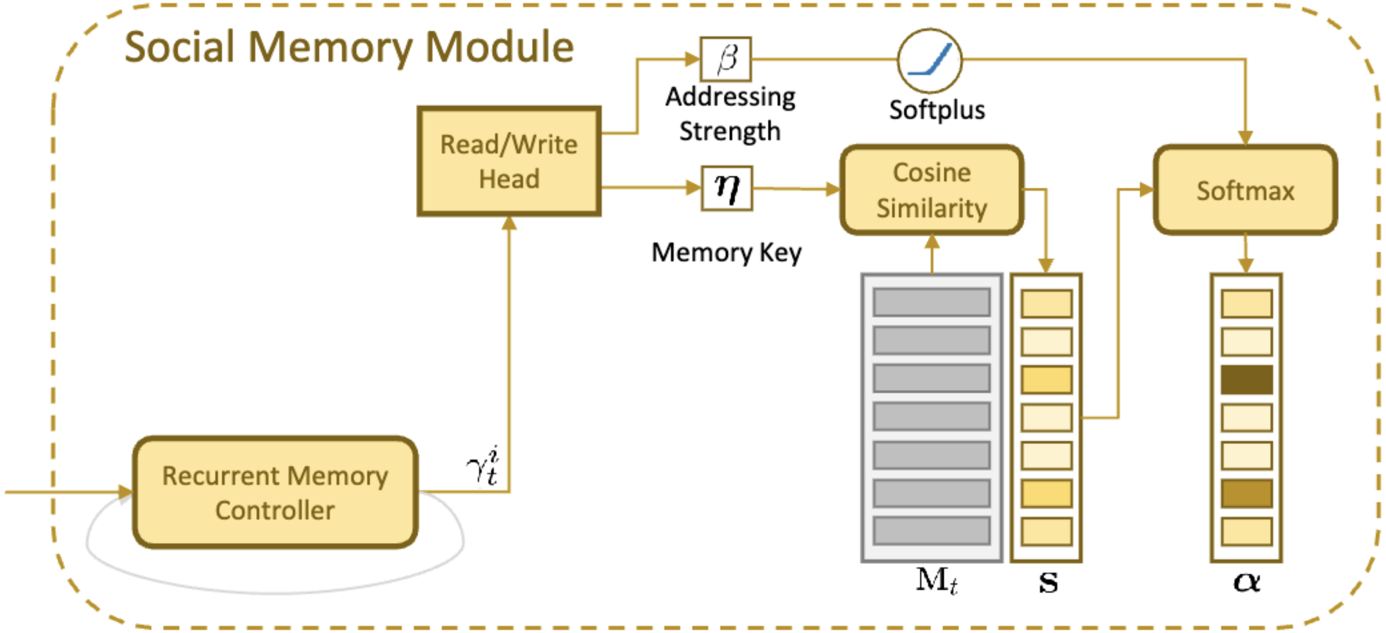


Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

Social Memory Addressing

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Read and Write steps share an *addressing* step
- Weights α identify the relevance of memory cells
- The controller outputs at each timestep a feature γ_t^i and feed it to R/W heads to get key η
- R/W head also generates a temperature β controlling the normalization of similarities $s_j = \frac{\eta m_j}{\|\eta\|\|m_j\|}$ in the softmax



Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

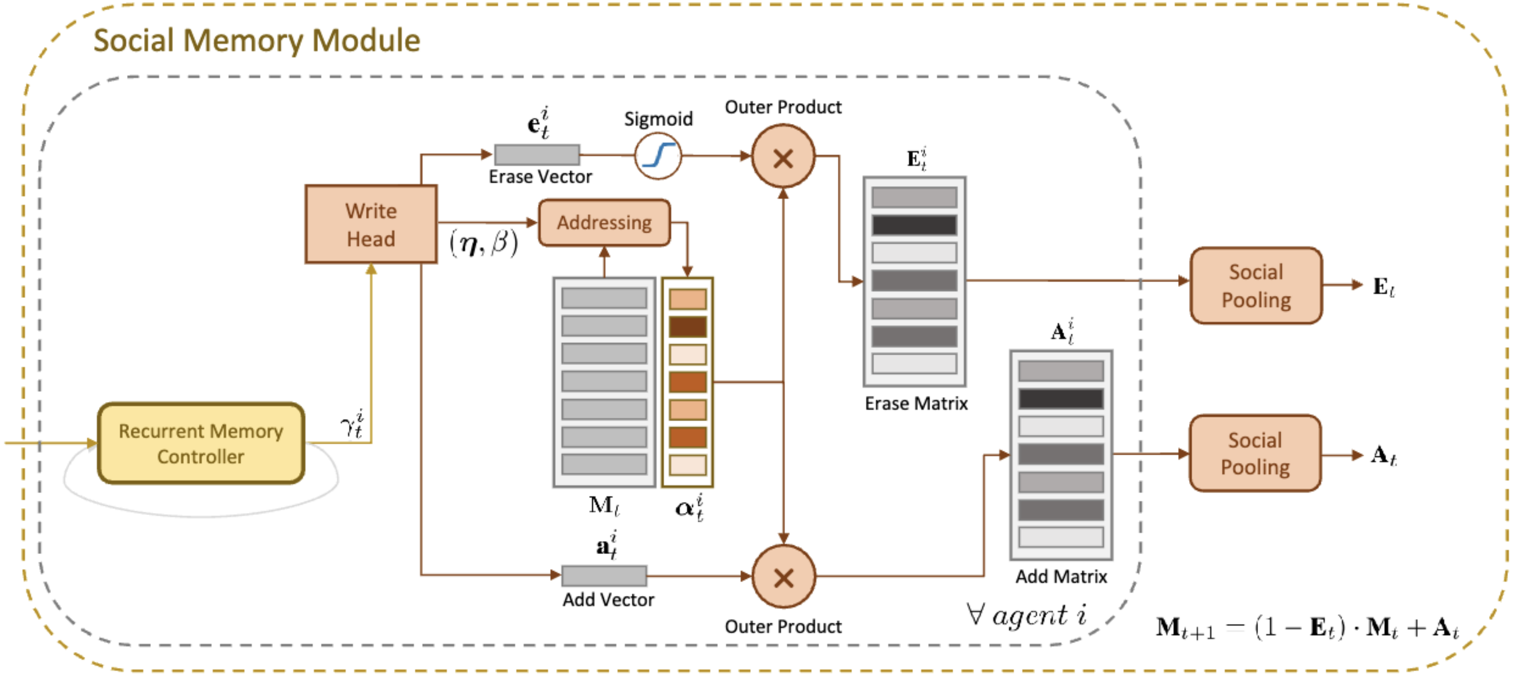
Social Memory Writing

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Write head will produce *erase* and *add* vectors
- Combining $\boldsymbol{\alpha}$ with vectors $\boldsymbol{e}_t^i$ and $\boldsymbol{a}_t^i$ using an outer product, we obtain *erase* and *add* matrices $\mathbf{E}_t, \mathbf{A}_t$
- Memory is updated every step

$$\mathbf{M}_{t+1} = (1 - \mathbf{E}_t) \cdot \mathbf{M}_t + \mathbf{A}_t$$

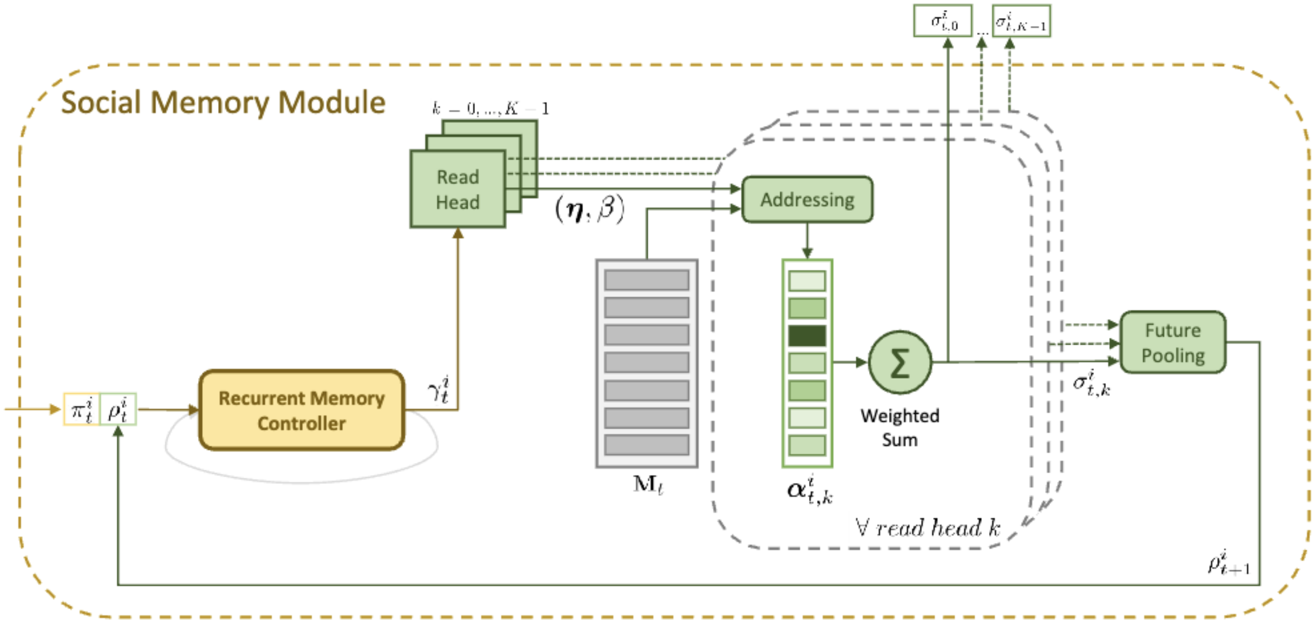
- We update after Social Pooling to be invariant to agent writing order



Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

Social Memory Reading

- For each agent i, separate read heads perform a memory addressing to obtain K social features $\sigma_{t,k}^{i}$
- $\sigma_{t,k}^{i}$ are fed in parallel into the decoder to generate a multimodal future prediction.
- The social features are then pooled together via *Future Pooling* and fed back to the model auto-regressively.



Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

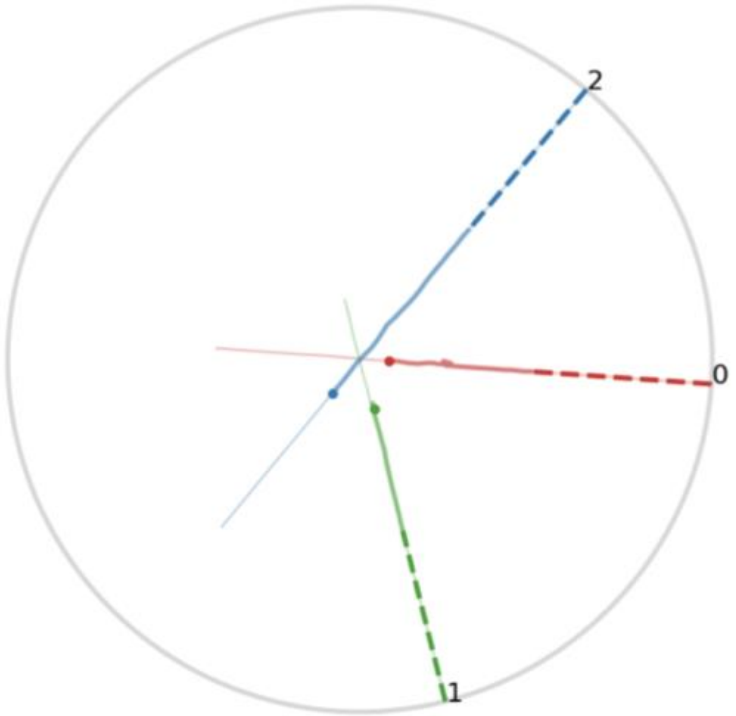
Results

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Synthetic Social Agents (SSA) a *"toy"* dataset with synthetic agents behaving according to a simple social rule

- Who gets to the center first has the right of way. Agents have random initial location and speed

Method	ADE ↓	FDE ↓	Kendall ↑
Linear	0.552 ±0.004	0.855 ±0.006	0.665 ±0.004
MLP	0.527 ±0.004	0.832 ±0.003	0.638 ±0.010
GRU ENC-DEC	0.525 ±0.004	0.829 ±0.003	0.642 ±0.009
Expert-Goals [35]	0.571 ±0.005	0.896 ±0.007	0.495 ±0.006
PECNet [34]	0.286 ±0.012	0.828 ±0.009	0.705 ±0.003
Trajectron++ [26]	0.519 ±0.011	0.818 ±0.019	0.569 ±0.015
Social-GAN [6]	0.302 ±0.004	0.506 ±0.003	0.626 ±0.031
AgentFormer [25]	0.243 ±0.003	0.385 ±0.003	0.701 ±0.006
SR-LSTM [66]	0.217 ±0.004	0.409 ±0.003	0.777 ±0.012
SMEMO	**0.169** ±0.006	**0.244** ±0.012	**0.827** ±0.008



Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

Results

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

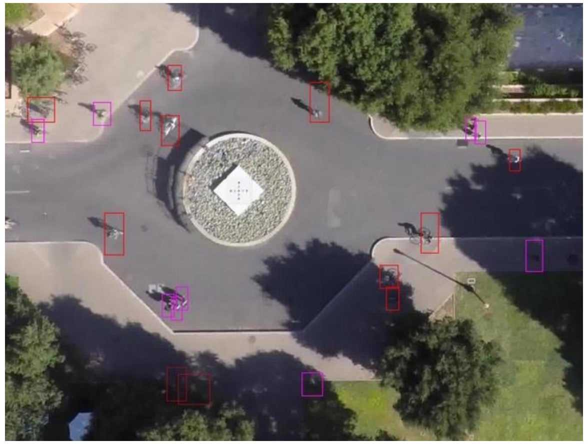
- Results on Stanford Drone with different settings
  (K = #futures)

K=20

Method	ADE	FDE	Method	ADE	FDE
Trajectron++ [26]*	19.30	32.70	MID [29]	9.73	15.32
SoPhie [8]	16.27	29.38	MANTRA [14]	8.96	17.76
EvolveGraph [67]	13.90	22.90	LB-EBM [70]	8.87	15.61
CF-VAE [71]	12.60	22.30	PCCSNet [69]	8.62	16.16
P2TIRL [72]	12.58	22.07	MemoNet [53]	8.56	12.66
Goal-GAN [33]	12.20	22.10	LeapFrog [28]	8.48	**11.66**
Expert-Goals [35]	10.49	13.21	Y-Net [36]	8.25	12.10
SimAug [73]	10.27	19.71	**SMEMO**	**8.11**	13.06
PECNet [34]	9.96	15.88			

K=5

Method	ADE	FDE
DESIRE [19]	19.25	34.05
Ridel et al. [68]	14.92	27.97
MANTRA [14]	13.51	27.34
PECNet [34]	12.79	25.98
PCCSNet [69]	12.54	-
TNT [32]	12.23	21.16
SMEMO	**11.64**	**21.12**

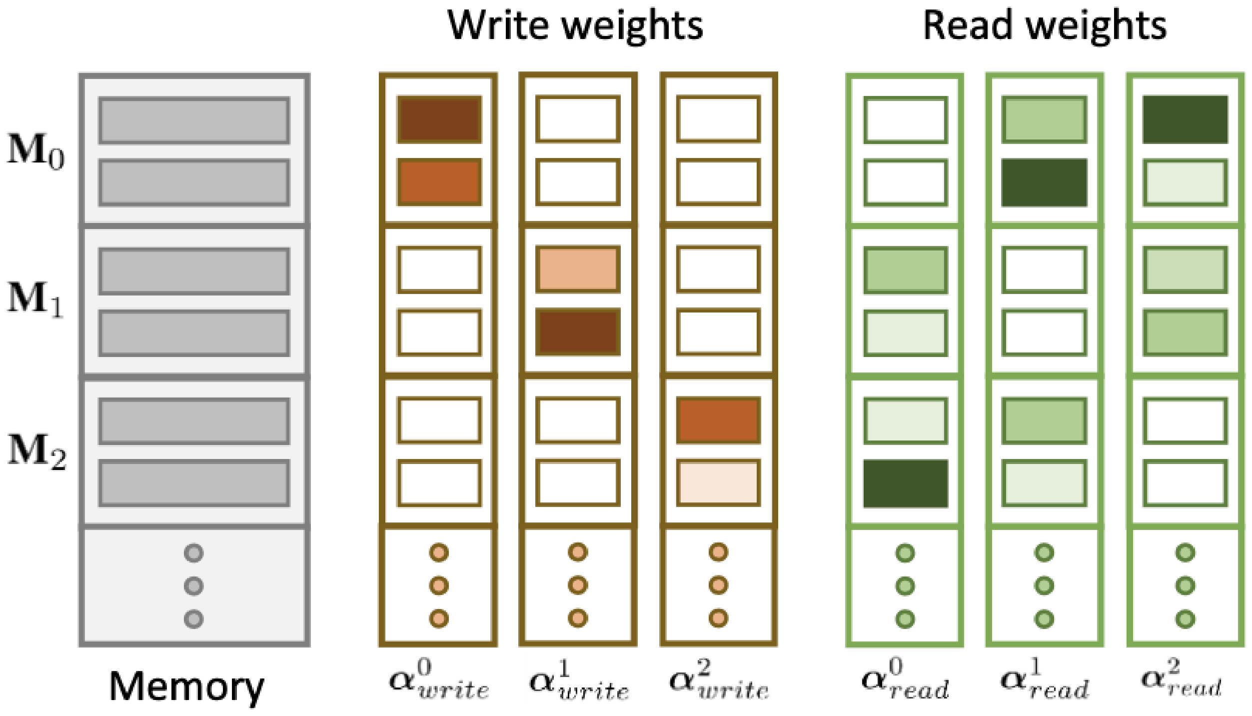


Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

Explainability via Social Memory

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Memory is partitioned in segments
- Each segment is reserved for a single agent
- R/W weights are actionable for explainability



Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

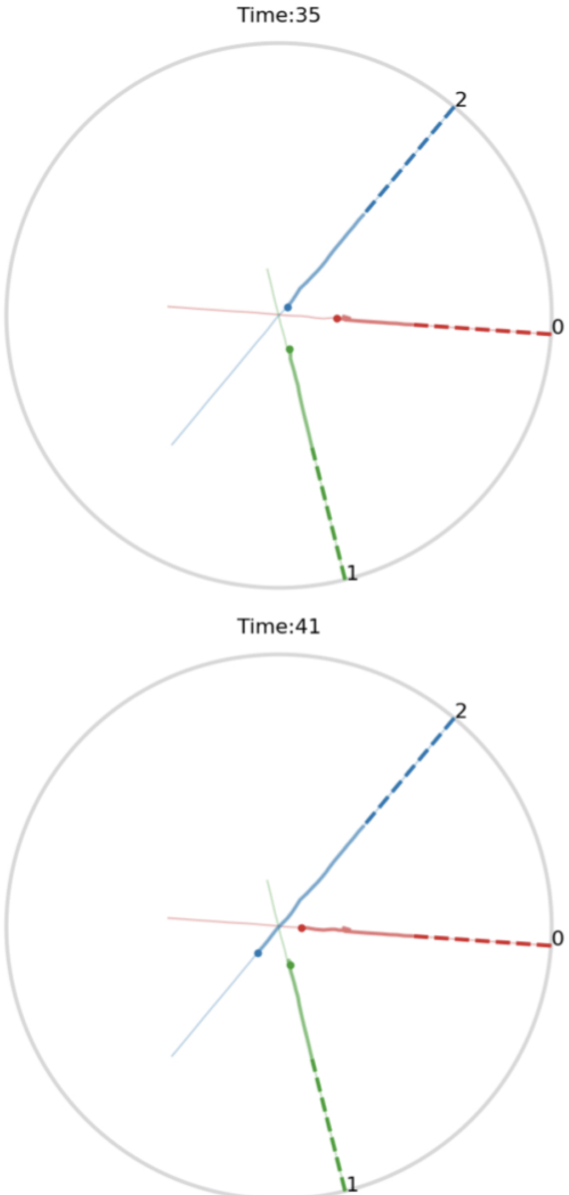
Explainability Results (ETH)

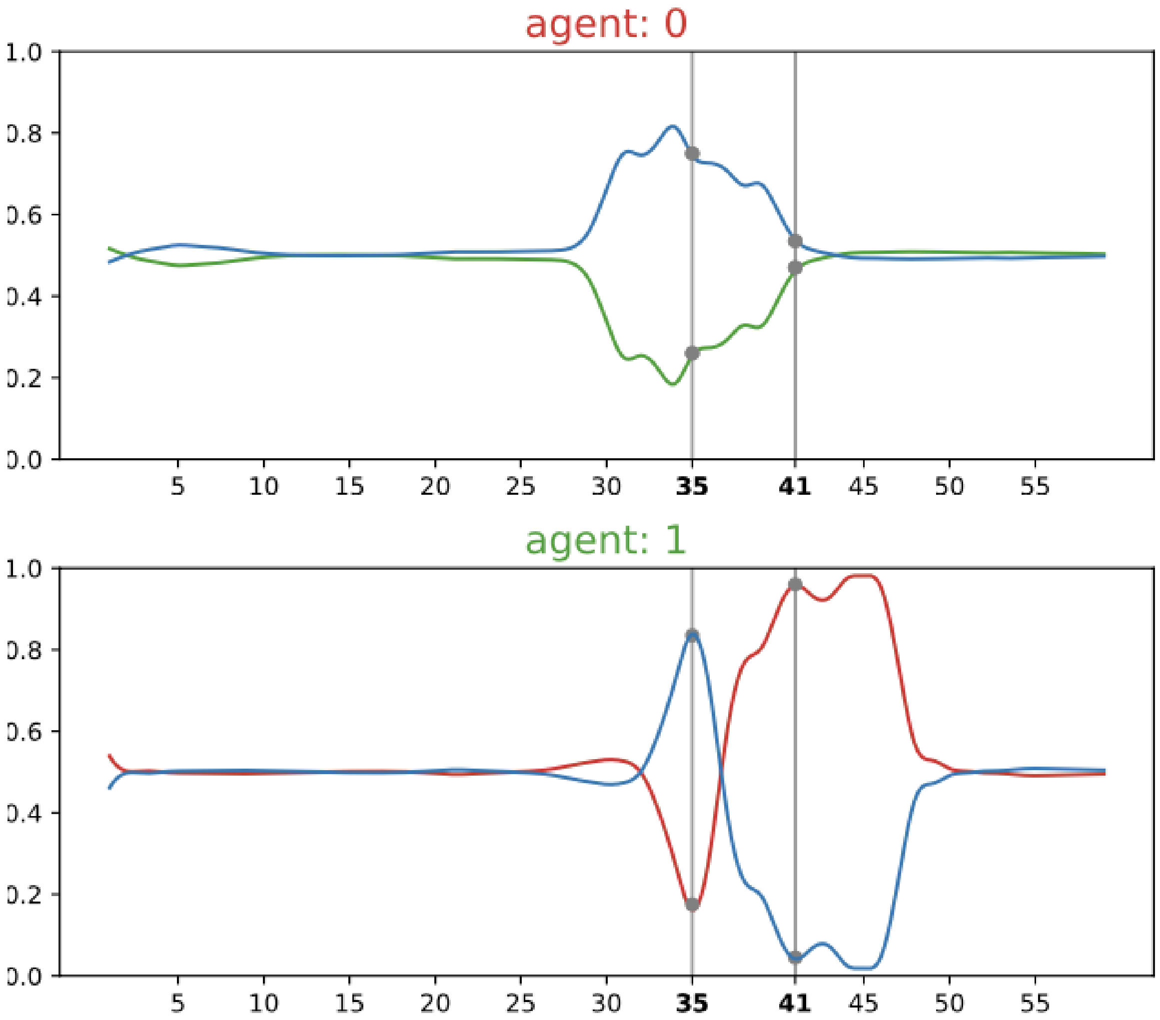
- Agent 0 and Agent 1 ignore Agent 2

Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

Explainability Results (SSA)

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE





A0 reads A2 = most likely to collide

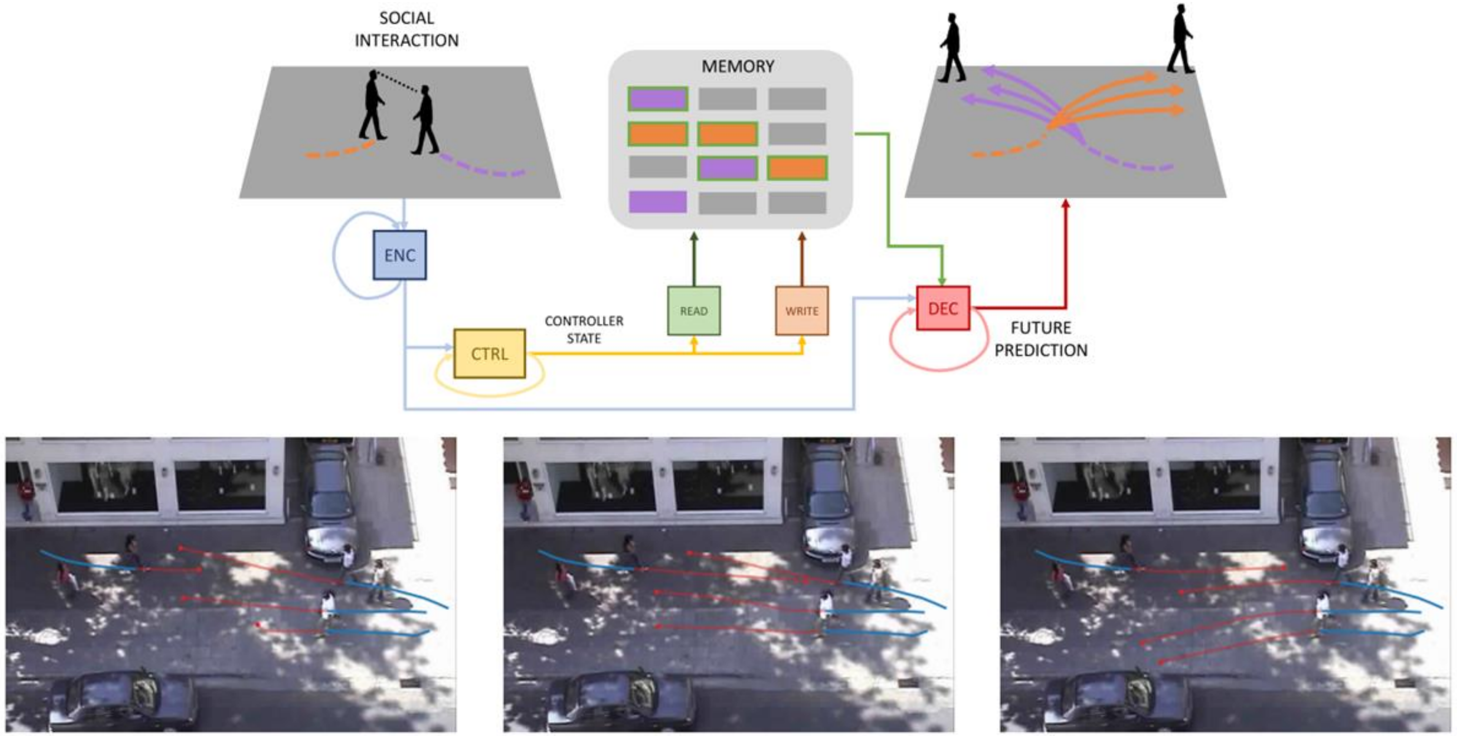
After A1 and A2 become equally (ir)relevant

A1 reads A2 = most likely to collide

After crossing attention switches

Code Available!

https://github.com/Marchetz/SMEMO_trajectory_forecasting



Social Multi-modal Prediction

Marchetti, F., Becattini, F., Seidenari, L. and Del Bimbo, A., 2024. SMEMO: social memory for trajectory forecasting. IEEE TPAMI.

Funding & Collaborators

- Work done in collaboration with



Dr. Francesco Marchetti   Dr. Federico Becattini   Prof. Alberto Del Bimbo

- Projects partially funded by

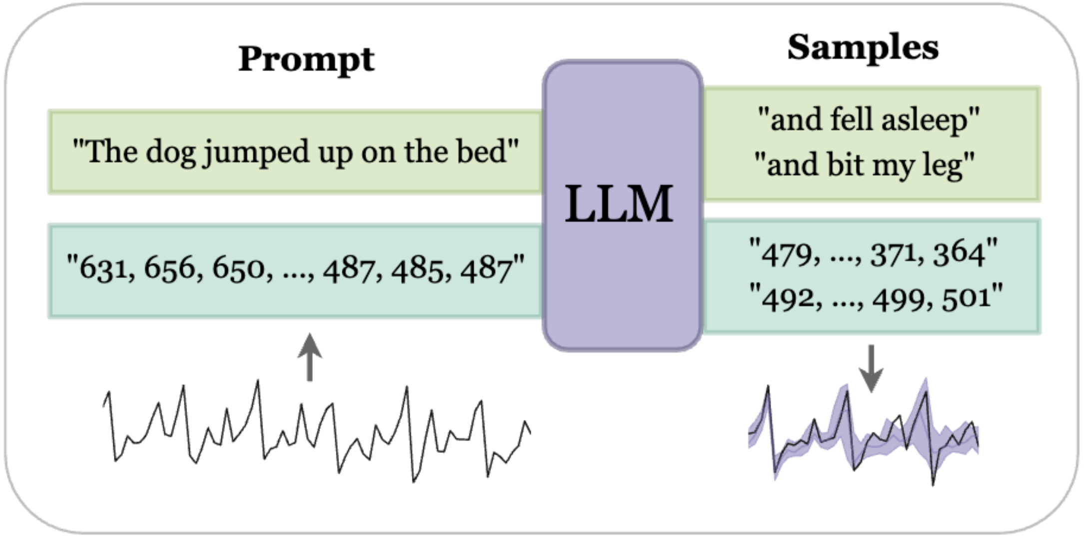
IMRA   AI4media

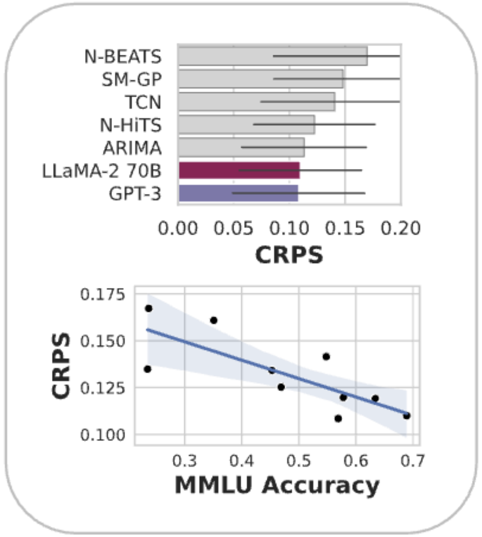
Foundation Models for Time Series

Foundation Models

LLM as Zero-Shot Learners

IDEA: Encode time series as text and prompt foundational LLM (GPT-4, LLaMA etc) to complete the sequence





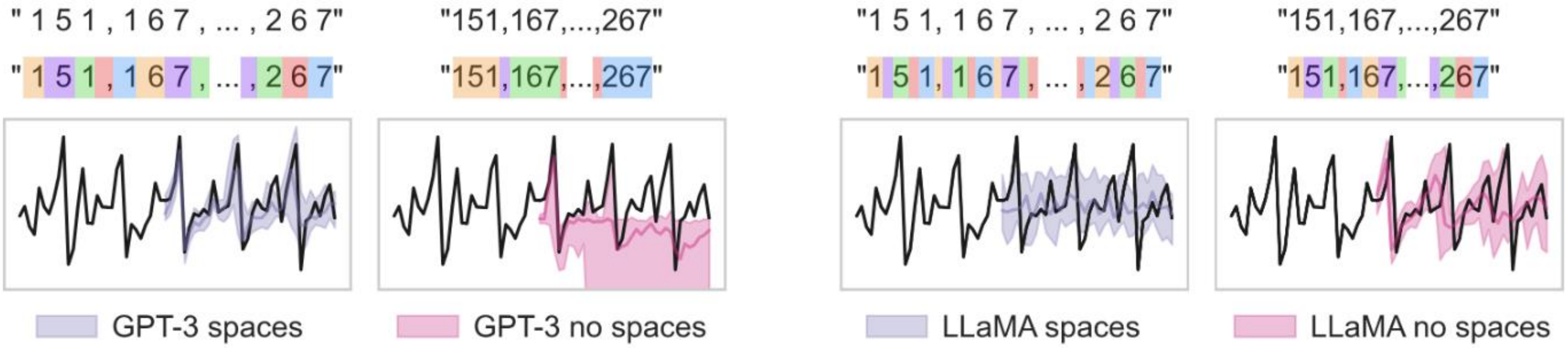
Gruver, Nate, et al. "Large language models are zero-shot time series forecasters." *Advances in Neural Information Processing Systems* 36 (2024).

Foundation Models

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

LLM as Zero-Shot Learners

CAVEAT: Tokenization is key!



Gruver, Nate, et al. "Large language models are zero-shot time series forecasters." *Advances in Neural Information Processing Systems* 36 (2024).

Foundation Models

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

LLM as Zero-Shot Learners

CAVEAT: Tokenization is key!

Scale values down so that the α-percentile of rescaled time series values is 1

Forecasting LLM can be sampled (adjusting T). When forecasting multiple estimates (20) are drawn and the median is used as point estimate

Representation Fixed precision is used with spaces to separate digits and commas values to separate values

$$0.123, 1.23, 12.3, 123.0 \rightarrow \text{" 1 2 , 1 2 3 , 1 2 3 0 , 1 2 3 0 0".}$$

Bonus missing values can be inserted as NaN (text)

$$[64, , , 49, , 16,] \rightarrow \text{"64, NaN, NaN, 49, NaN, 16, NaN"}$$

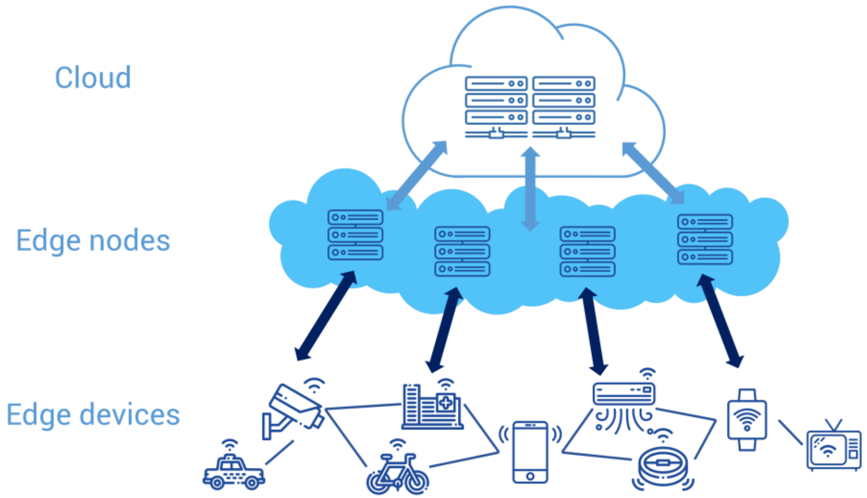
Gruver, Nate, et al. "Large language models are zero-shot time series forecasters." *Advances in Neural Information Processing Systems* 36 (2024).

Case Study: IoE Networks

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Edge computing is a promising solution for enabling pervasive Internet of Everything (IoE) environments, connecting all objects for intelligent, distributed systems across hybrid domains.

Edge computing enables data processing near the source, reducing latency and bandwidth. AI integration in these networks ensures quick adaptation, reliable connectivity, and flexibility in managing diverse traffic in hybrid environments.

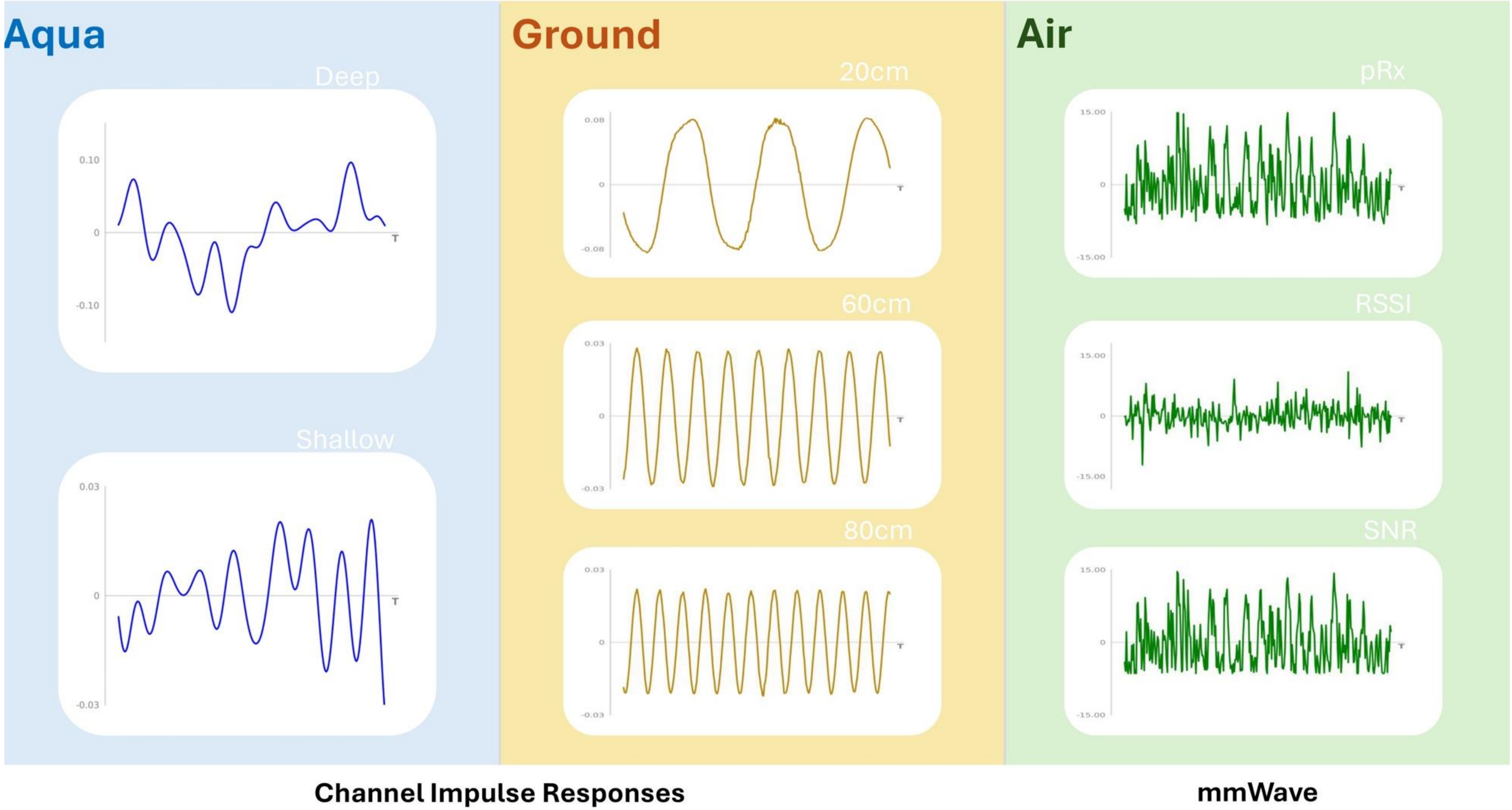


Predicting communication channel conditions quickly and accurately is crucial for ensuring quality service in AI-IoE networks. Current AI solutions require large datasets and frequent retraining, which are costly and inefficient.

Data: channel state

IDEA: foundation model-based framework that enables the same edge node to interact with IoE networks deployed in different environments (i.e Aqua, Ground, Air).

Different dynamics and frequency between different environments.



Foundation Models for Time Series

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

The rise of foundation models in NLP has led to the development of specialized models for time-series data.

CHRONOS

Chronos[1] (by Amazon) is a pretrained probabilistic time series model that tokenizes scaled and quantized values, generating future values autoregressively using a T5-based **encoder-decoder architecture**

TimesFM

TimesFM[2] (by Google) is a **decoder-only** foundation model for time-series forecasting that splits data into patches as tokens and predicts the next patch in an **autoregressive** manner.

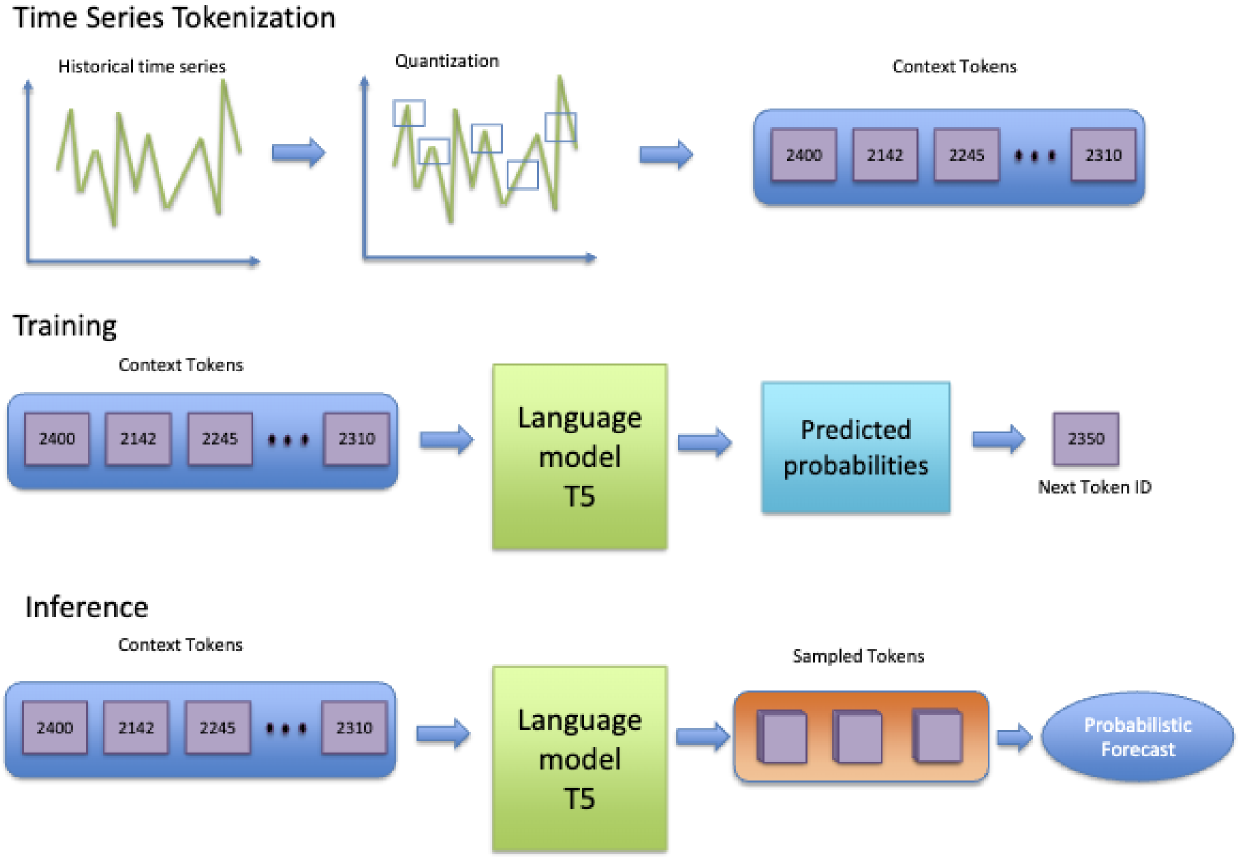
[1] Chronos: Learning the Language of Time Series, A. F. Ansari et al., arxiv

[2] A decoder-only foundation model for time-series forecasting, A. Das et al. n Proceedings of the 41st International Conference on Machine Learning, 2024

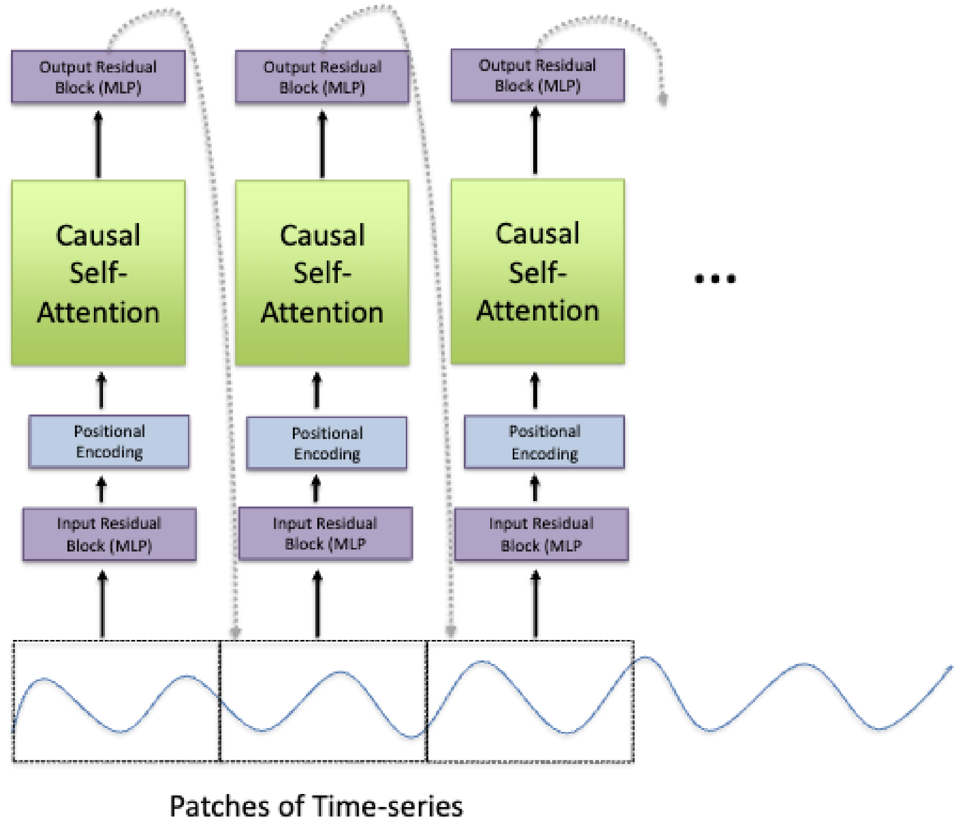
Foundation Models for Time Series

Architectural details of the two models

Chronos



TimesFM

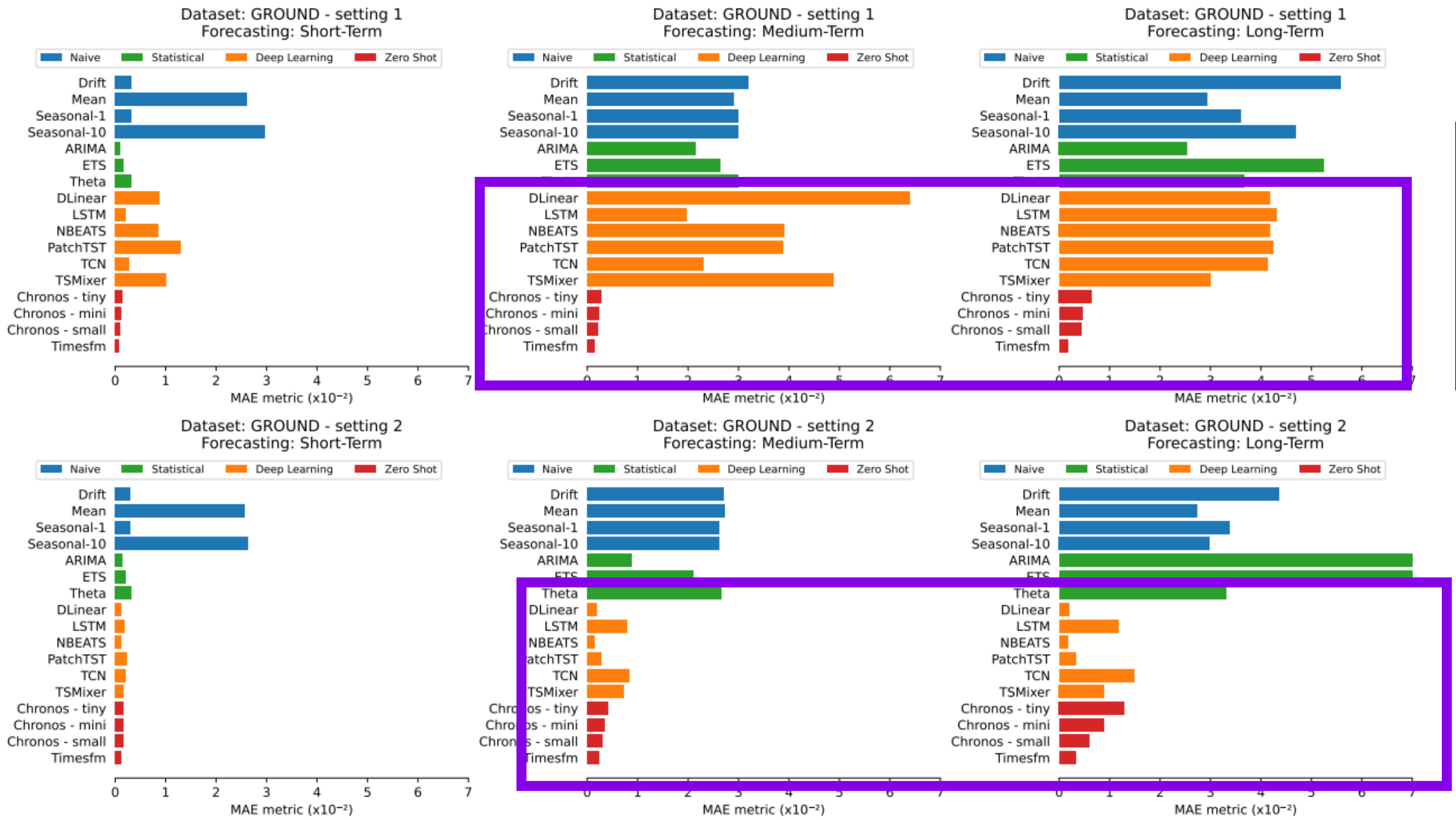


[1] Chronos: Learning the Language of Time Series, A. F. Ansari et al., arxiv

[2] A decoder-only foundation model for time-series forecasting, A. Das et al. n Proceedings of the 41st International Conference on Machine Learning, 2024

Results

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



Setting 1

Split data depending on receiver distance

ZS >> DL >> ALL

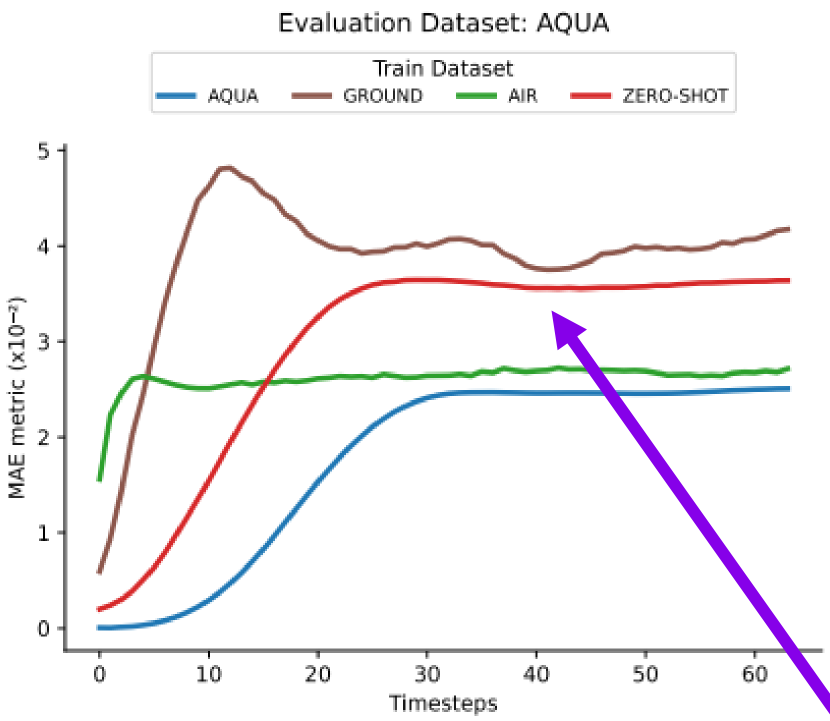
Setting 2

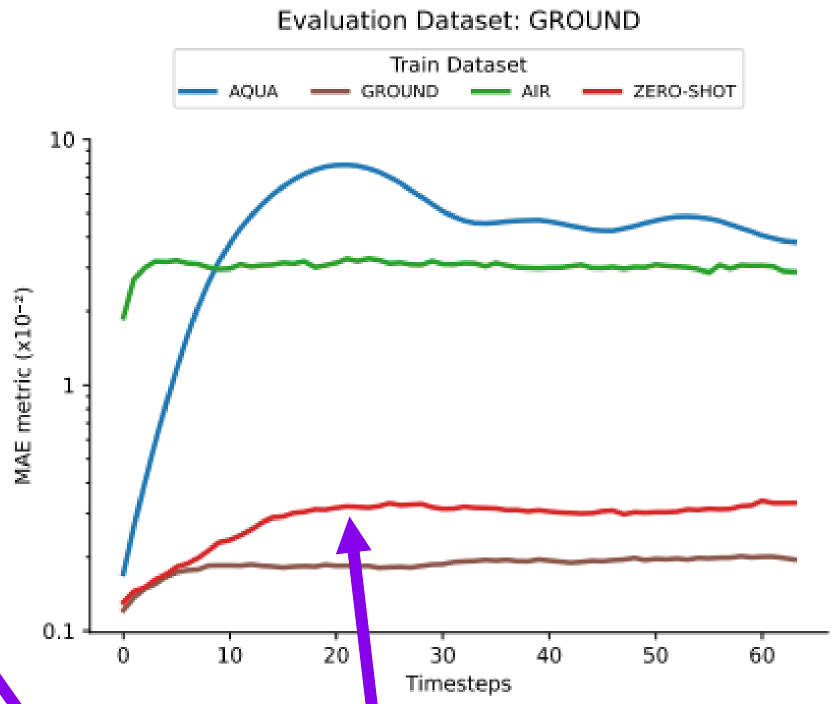
All distances present in train/val/test split

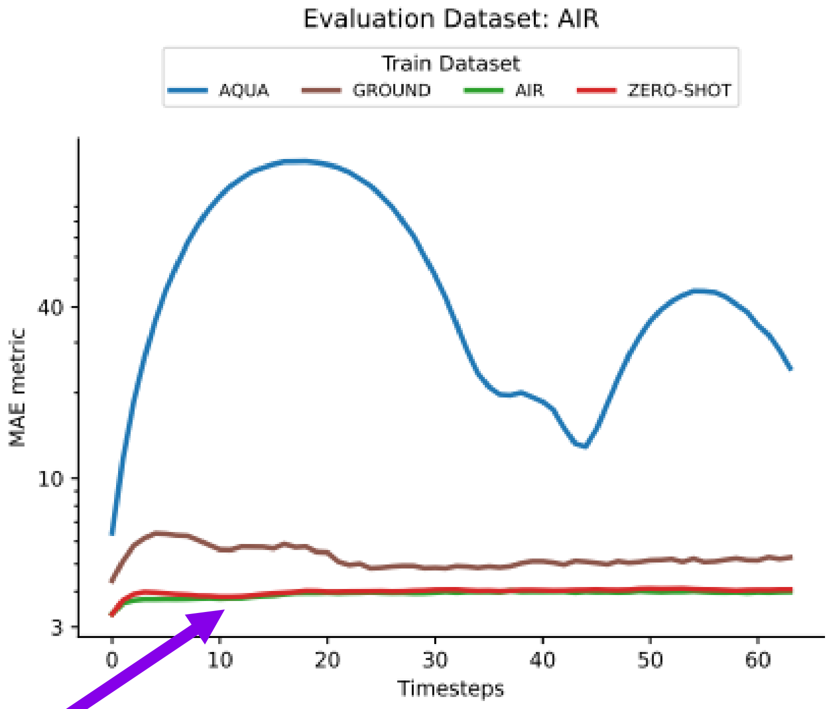
DL >> ZS >> ALL

Cross-Dataset Evaluation

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE







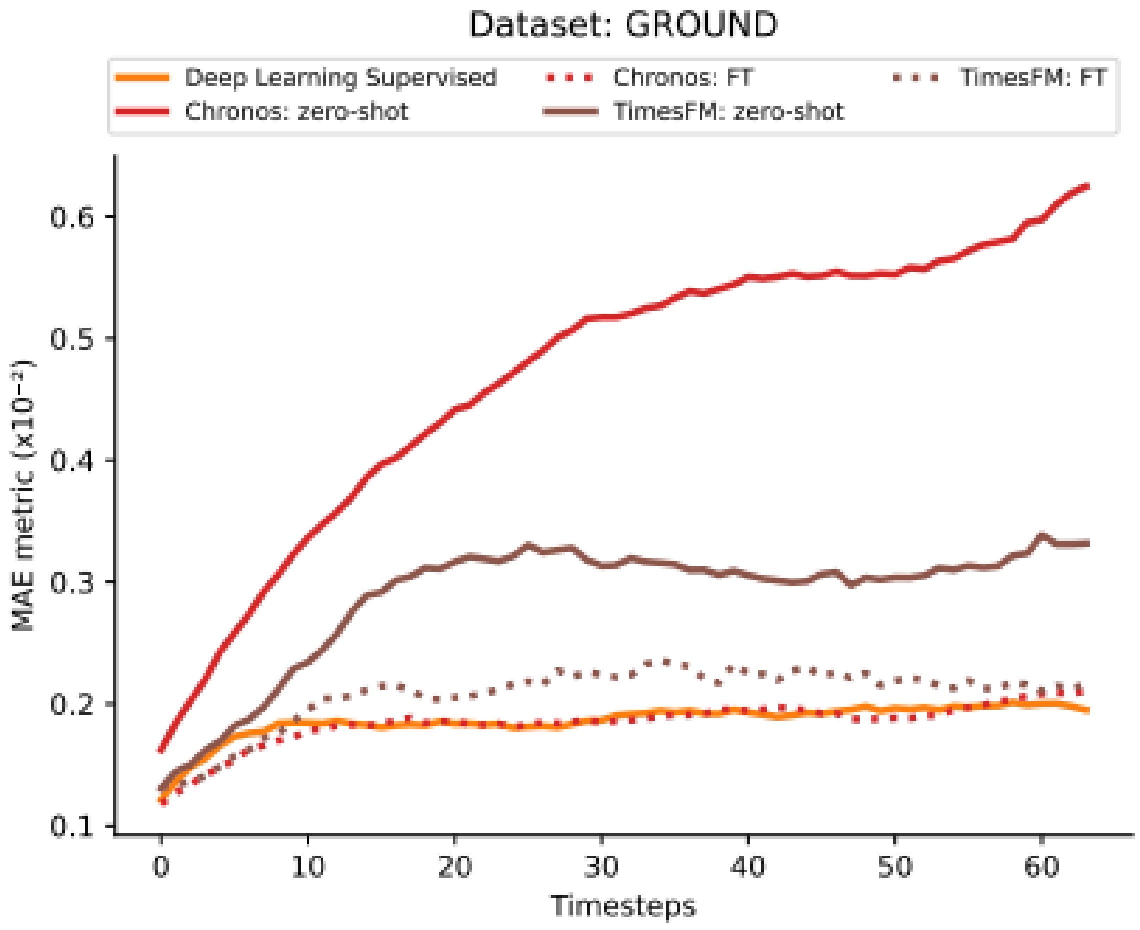
ZS gets competitive results vs Supervised Deep Learning Methods

Fine-tuning

The fine-tuned models attain performance on par with the supervised model on the specific dataset.

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

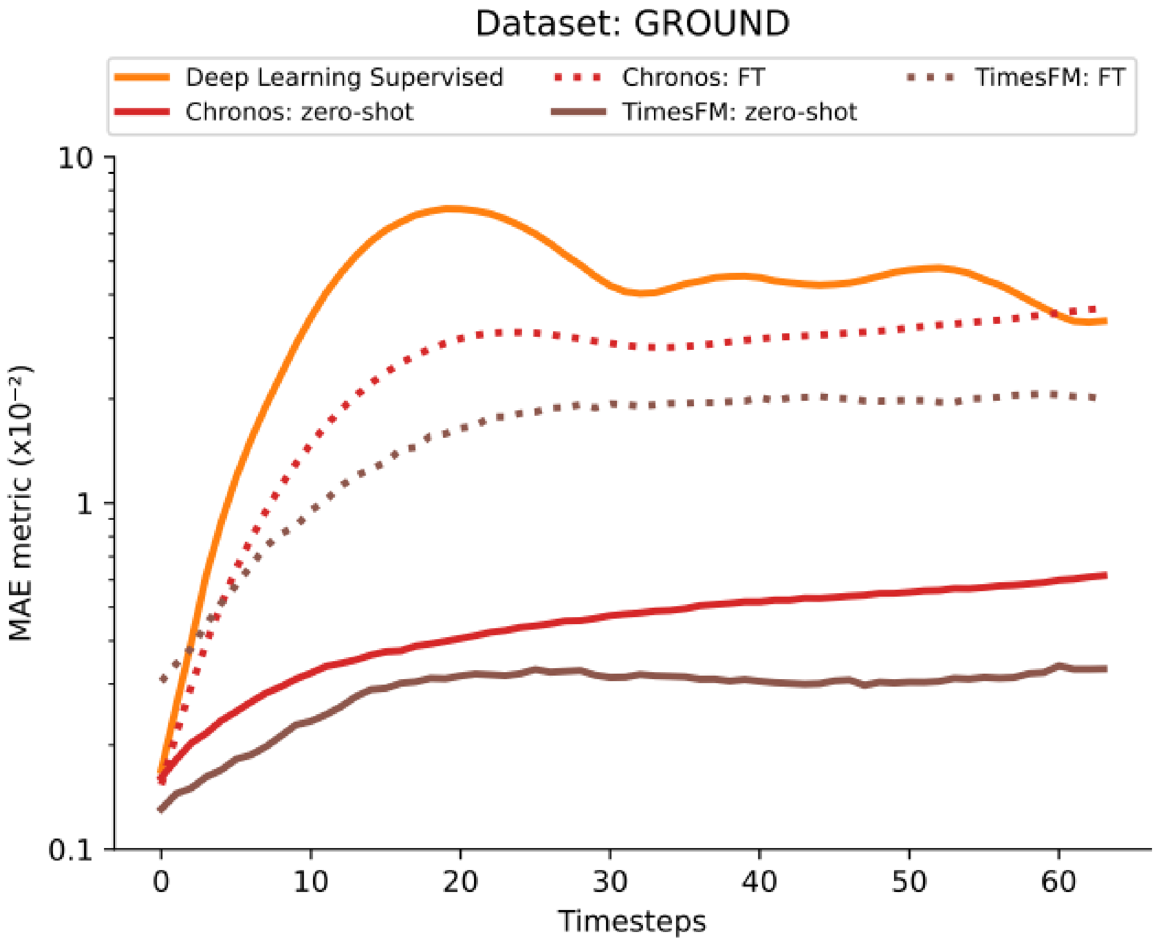
Fine-tune on GROUND
Test on GROUND



Fine-tuning

Fine-tuned foundation models maintain performance better on new datasets than models trained from scratch.

Fine-tune on AQUA
Test on GROUND



Funding & Collaborators

- Work done in collaboration with

Dr. Francesco Marchetti Dr. Benedetta Picano Prof. Romano Fantacci

- Projects partially funded by

Finanziato dall'Unione europea
NextGenerationEU

Conclusion

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Transformers are great a long-interaction and feature fusion. Still recurrent architectures provide lean yet powerful models
- Memory is an easy-to-plug block to allow more complex states, modelling continual learning, communication between agents etc…)
- Foundation Models are now available for I, V, L , I+L, V+L, T…
- What about World Foundation Models (JEPA, Genie etc)? Can they help in real-world quantities forecasting?

References

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

- Marchetti, Francesco, et al. "Mantra: Memory augmented networks for multiple trajectory prediction." **CVPR 2020**
- Marchetti, Francesco, et al. "Multiple trajectory prediction of moving agents with memory augmented networks." **IEEE TPAMI 2020**
- Marchetti, Francesco, et al. "CrossFeat: Semantic Cross-modal Attention for Pedestrian Behavior Forecasting." **IEEE TIV 2024.**
- Ciamarra Andrea, et al. FLODCAST: Flow and Depth Forecasting via Multimodal Recurrent Architectures. **Elsevier Pattern Recognition 2024**
- Marchetti, Francesco, et al. "SMEMO: Social Memory for Trajectory Forecasting." *arXiv preprint* (2024). **TPAMI 2024**
- Marchetti, Francesco, et al. "Foundation Forecasting in IoE Networks: When Generative AI Meets Programmable Edge Nodes" **IEEE IoT 2025** (under minor revision)

Questions?

lorenzo.seidenari@unifi.it

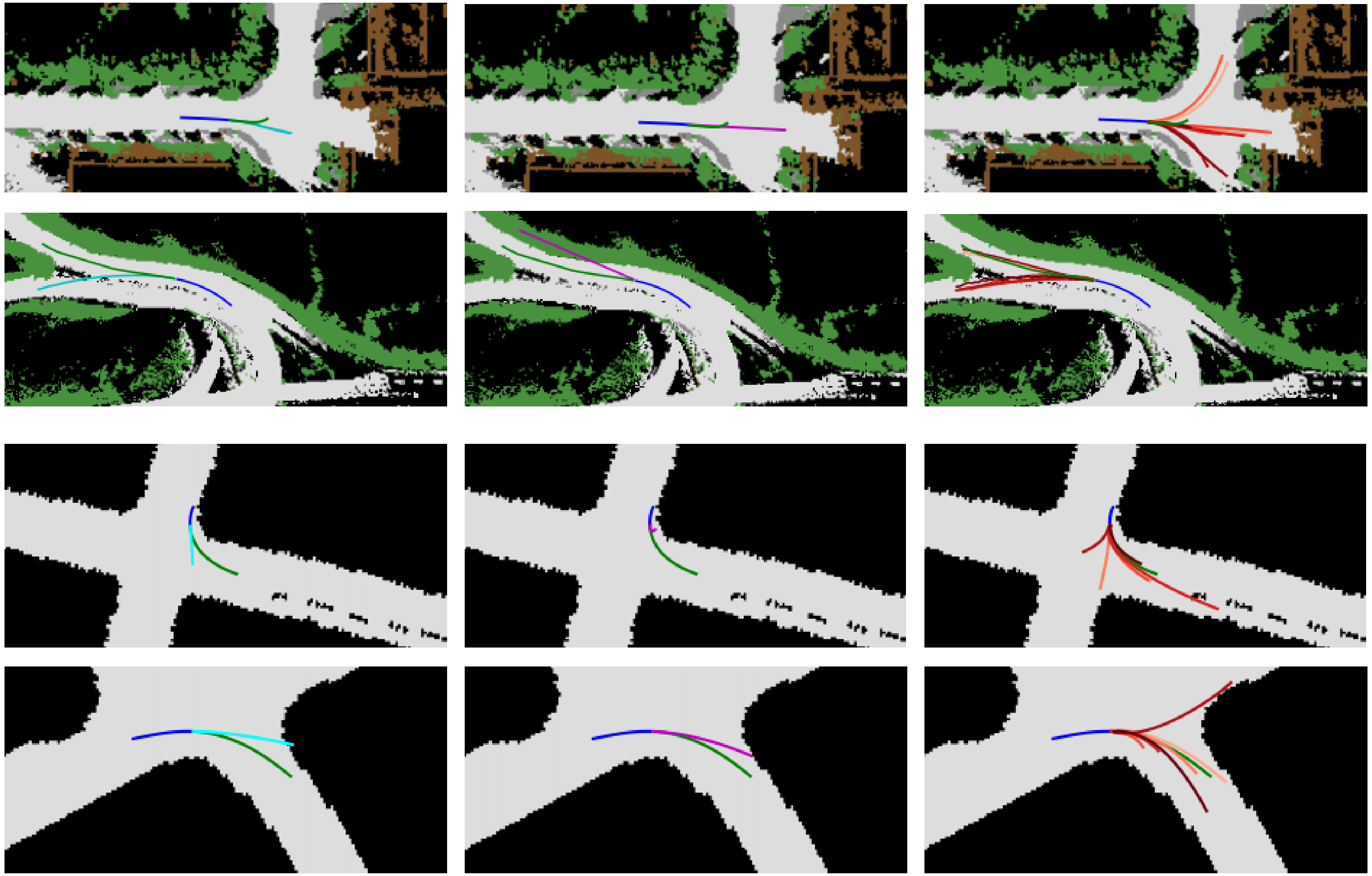
Extra

Qualitative Results

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

KITTI

ARGOVERSE



(a) Linear (b) Kalman (c) MANTRA

Quantitative Results

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Method	ADE				FDE			
	1s	2s	3s	4s	1s	2s	3s	4s
Kalman	0.51	1.14	1.99	3.03	0.97	2.54	4.71	7.41
Linear	0.20	0.49	0.96	1.64	0.40	1.18	2.56	4.73
MLP	0.20	0.49	0.93	1.53	0.40	1.17	2.39	4.12
MANTRA (top 1)	0.24	0.57	1.08	1.78	0.44	1.34	2.79	4.83
MANTRA (top 5)	0.17	0.36	0.61	0.94	0.30	0.75	1.43	2.48
MANTRA (top 10)	0.16	0.30	0.48	0.73	0.26	0.59	1.07	1.88
MANTRA (top 20)	**0.16**	**0.27**	**0.40**	**0.59**	**0.25**	**0.49**	**0.83**	**1.49**
DESIRE (top 1) [22]	-	-	-	-	0.51	1.44	2.76	4.45
DESIRE (top 5) [22])	-	-	-	-	0.28	0.67	1.22	2.06
DESIRE (top 20) [22])	-	-	-	-	-	-	-	2.04

Table 1. Results on the KITTI dataset. Results obtained by DESIRE are given as reference even if not comparable, due to the data collection process.

Method	ADE				FDE			
	1s	2s	3s	4s	1s	2s	3s	4s
Kalman	0.33	0.54	0.93	1.4	0.46	1.18	2.18	3.32
Linear	0.31	0.56	0.89	1.28	0.47	1.13	1.94	2.87
MLP	0.30	0.54	0.88	1.28	0.46	1.12	1.94	2.88
RNN Enc-Dec [38]	0.68	1.94	3.20	4.46	-	-	-	-
Markov [35]	0.70	1.41	2.12	2.99	-	-	-	-
Conv-LSTM (top 5) [35]	0.76	1.23	1.60	1.96	-	-	-	-
INFER (top 1) [35]	0.75	0.95	1.13	1.42	1.01	1.26	1.76	2.67
INFER (top 5) [35]	0.56	0.75	0.93	1.22	0.81	1.08	1.55	2.46
MANTRA (top 1)	0.37	0.67	1.07	1.55	0.60	1.33	2.32	3.50
MANTRA (top 5)	0.33	0.48	0.66	0.90	0.45	0.78	1.22	2.03
MANTRA (top 10)	0.31	0.43	0.57	0.78	0.43	0.67	1.04	1.78
MANTRA (top 20)	**0.29**	**0.41**	**0.55**	**0.74**	**0.41**	**0.64**	**1.00**	**1.68**

Table 2. Results on the KITTI dataset (INFER split).

Method	ADE				FDE			
	1s	2s	3s	4s	1s	2s	3s	4s
INFER (top 1) [35]	1.06	1.35	1.48	1.68	1.31	1.71	1.70	2.56
INFER (top 5) [35]	0.85	1.14	1.29	1.50	1.18	1.58	1.58	2.41
MANTRA (top 1)	0.55	0.77	1.01	1.30	0.60	1.15	1.82	2.63
MANTRA (top 5)	0.55	0.68	0.82	1.03	0.58	0.88	1.37	2.07
MANTRA (top 10)	0.44	0.56	0.72	0.94	0.48	0.73	1.33	1.98
MANTRA (top 20)	**0.31**	**0.43**	**0.59**	**0.83**	**0.35**	**0.61**	**1.24**	**1.96**

Table 3. Results on the Oxford RobotCar dataset.

Method	ADE	FDE
Conv-LSTM (top 1) [35]	1.50	-
Conv-LSTM (top 3) [35]	1.36	-
Conv-LSTM (top 5) [35]	1.28	-
INFER (top 1) [35]	1.11	1.59
INFER (top 3) [35]	0.99	1.45
INFER (top 5) [35]	0.91	1.38
MANTRA (top 1)	0.81	1.42
MANTRA (top 3)	0.66	1.15
MANTRA (top 5)	0.60	1.00
MANTRA (top 10)	0.54	0.86
MANTRA (top 20)	**0.49**	**0.79**

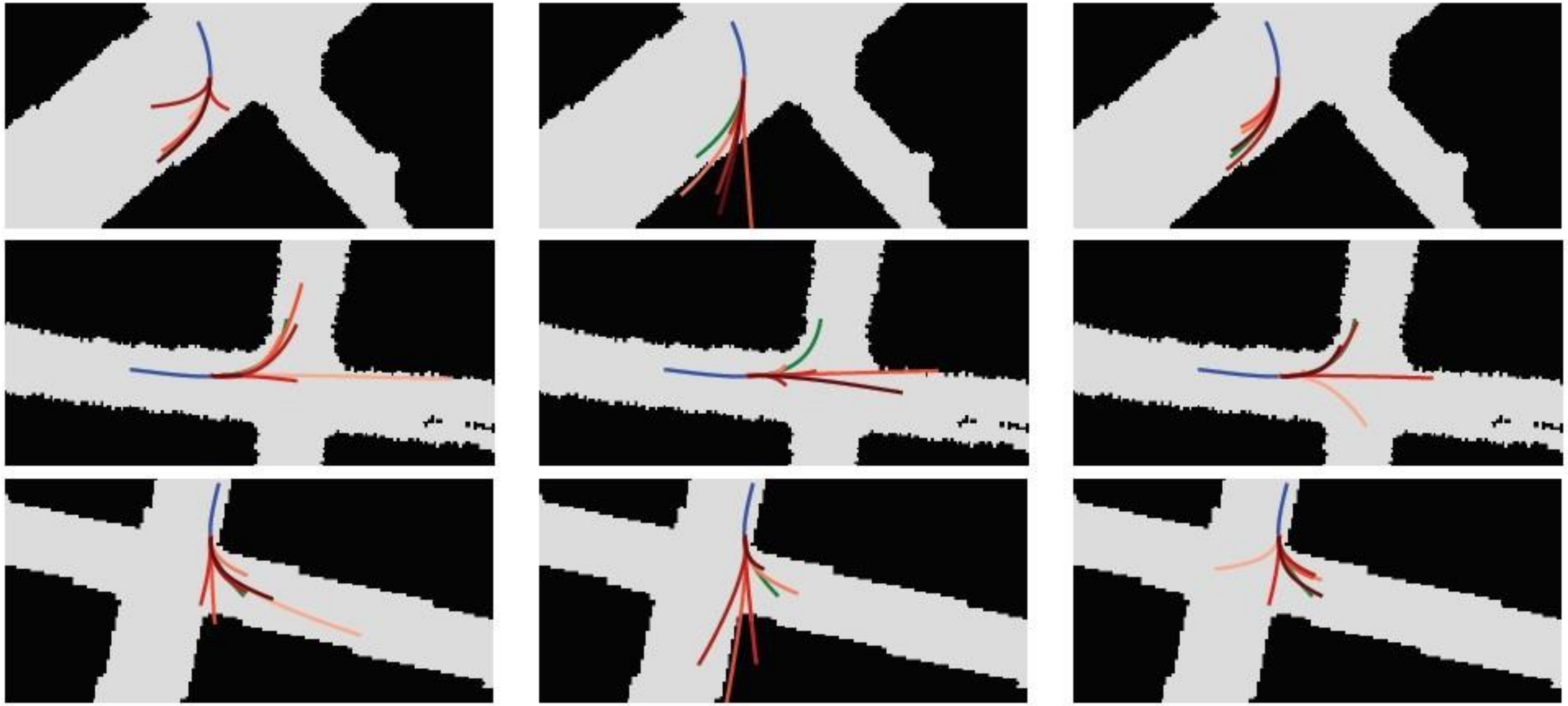
Table 4. Results on the Cityscapes dataset at 1s in the future.

Ablation Studies

Method	ADE		FDE		Off-Road (%)	Memory Size
	1s	3s	1s	4s		
MANTRA (Full)	0.56	**1.22**	0.84	**2.30**	3.15	6397 (3.1 %)
MANTRA (Controller Past)	**0.52**	**1.22**	**0.79**	2.38	8.14	**6242 (2.9 %)**
MANTRA (Controller Context)	0.73	1.87	1.19	3.70	**3.08**	21992 (10.5 %)
MANTRA w/o dec.	0.80	1.47	1.12	2.44	6.01	6397 (3.1 %)
MANTRA w/o rot. inv.	1.11	2.54	1.80	4.63	40.26	75674 (36.3 %)

Ablation Studies

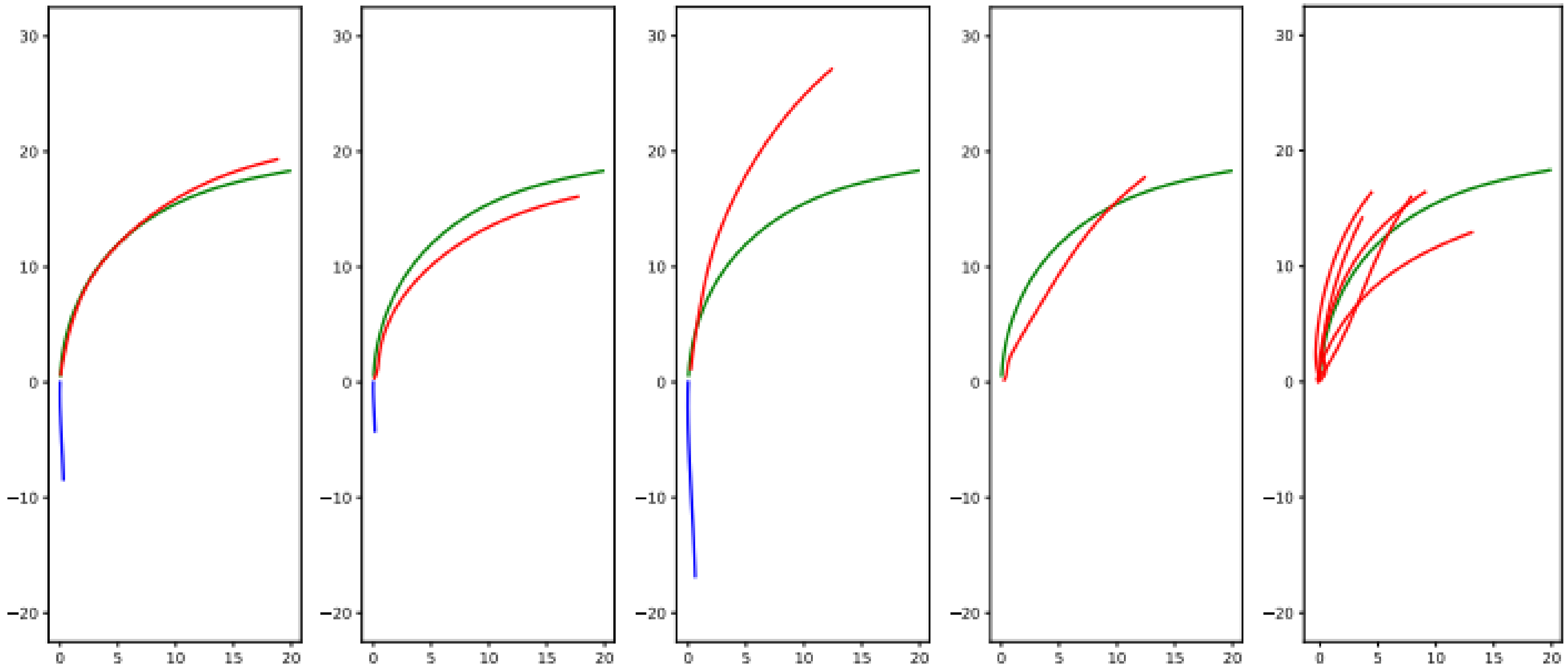
UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



(a) Controller Context (b) Controller Past (c) Full

Decoder Analysis: past

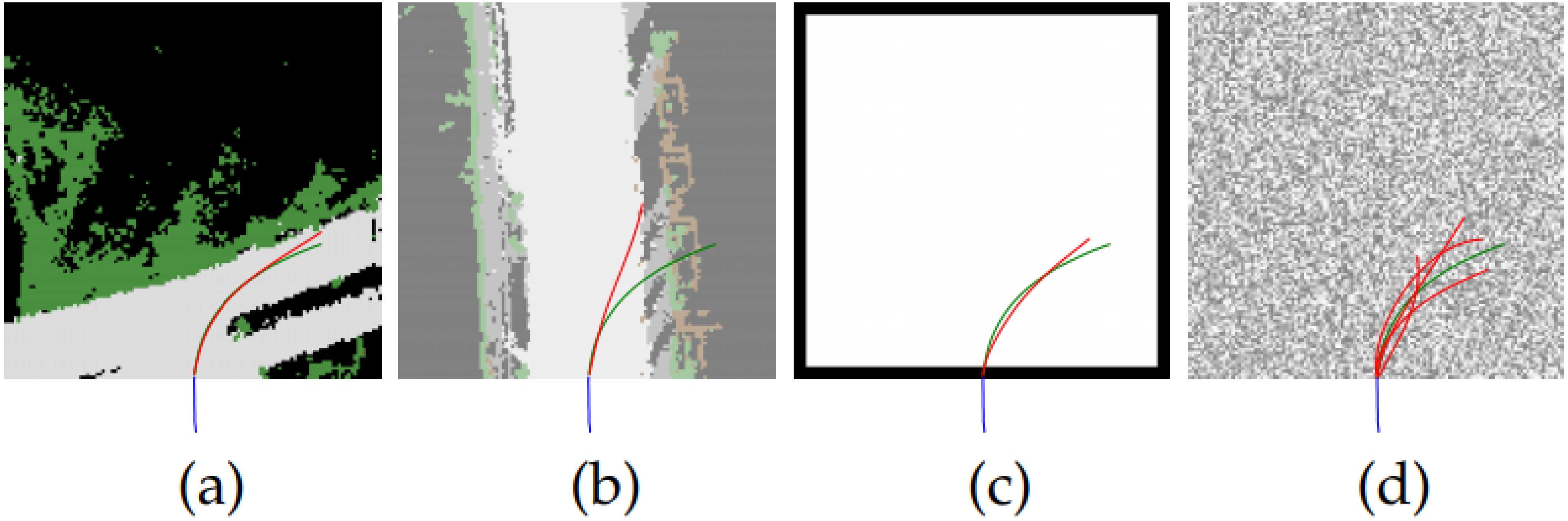
UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE



(a)	(b)	(c)	(d)	(e)
observed past	slower past	faster past	zeroed past	randomized pasts

Decoder Analysis: context

UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE



CONTEXT:	Original	different	Embedding zeroed	Multiple randomized embeddings

Reading Controller weights

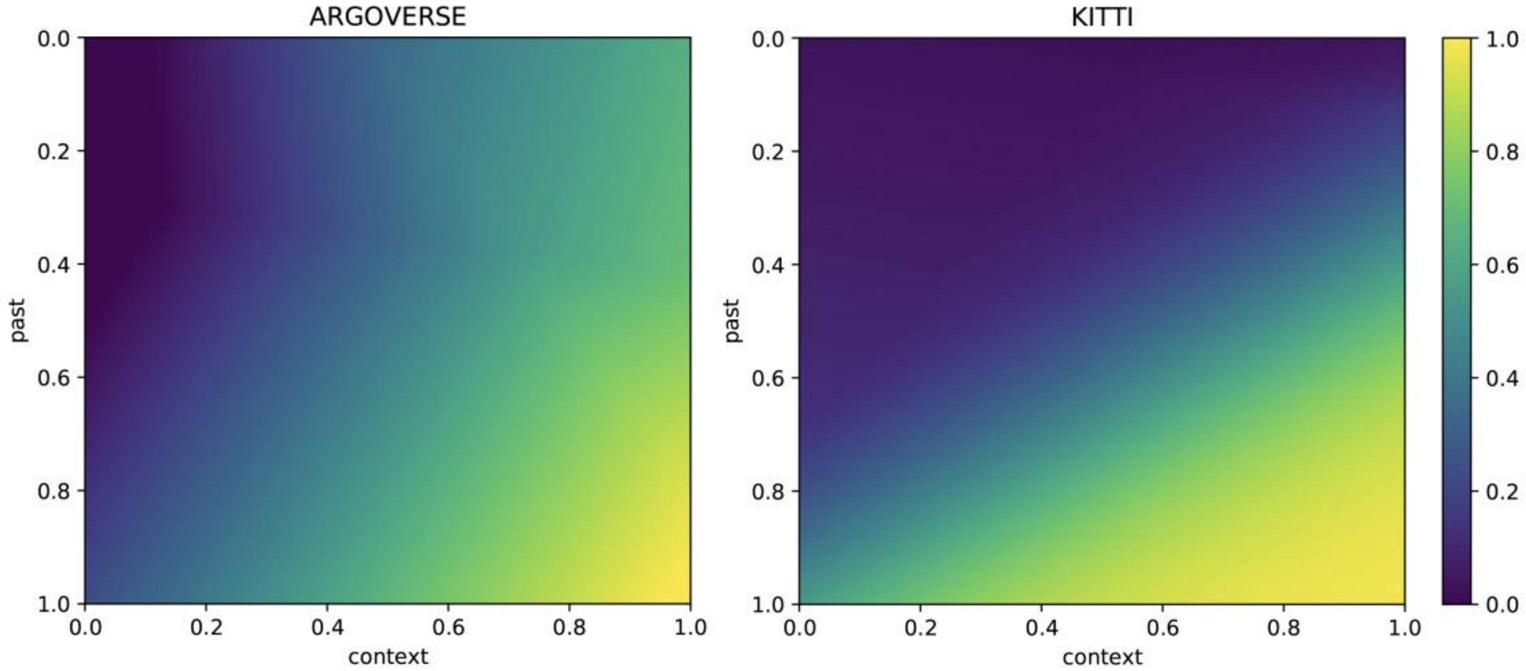


Fig. 8. Reading controller scores varying past and context similarities. Different blending functions are learned for different datasets, privileging the past on KITTI and increasing the relevance of context on Argoverse.

Memory Inspection

UNIVERSITÀ
DEGLI STUDI
FIRENZE
DINFO
DIPARTIMENTO DI
INGEGNERIA
DELL'INFORMAZIONE

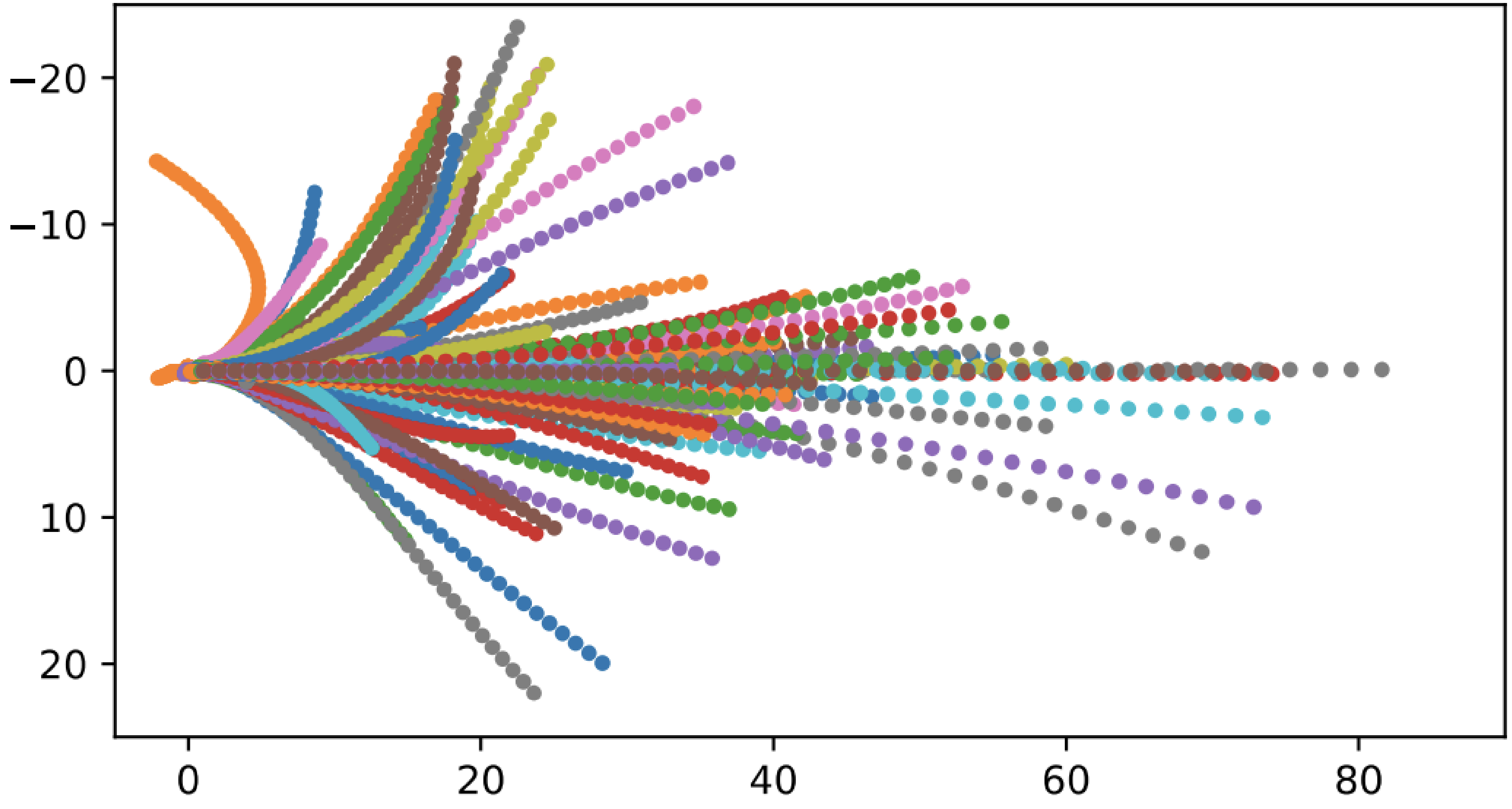


Fig. 9. Decoded trajectories from memory.

Data: channel state

Aqua Domain Dataset
Measurements of long-range underwater acoustic channel impulse responses:
- **Deep environment**: 100 km distance, 1800 m water depth
- **Shallow environment**: 50 km distance, 60 m water depth

Ground dataset
The channel impulse responses were detected inside an anechoic chamber with five different distances between transmitter and receivers (20cm, 30cm, 40cm, 60cm, 80cm).

Air Domain Dataset
UAV communications in mmWave spectrum. Key metrics: RSSI, received power (pRx), and SNR.

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE