Emergent self-awareness in multi-sensor physical agents

Joint Doctorate in Interactive and Cognitive Environments

Giulia Slavic

UNIGE Supervisors: Prof. Carlo Regazzoni, Prof. Lucio Marcenaro

UC3M Supervisor: Prof. David Martín Gómez

UNIVERSITÀ DEGLI STUDI DI GENOVA

Table of Contents

Research Objective and Motivation

Research Objective

Research Title:

Emergent self-awareness in multi-sensor physical agents

Self-awareness (SA) + Awareness:
knowledge of state and surroundings.

Emergent: knowledge acquired in an unsupervised way.

The agent learns what is new in unseen situations.

Development of self-aware models for autonomous vehicles that leverage the combination of multiple sensors. Focus is given to the video sensor.

Autonomous Vehicles

Vehicles designed to diminish or eliminate the need for human intervention in the execution of their tasks.

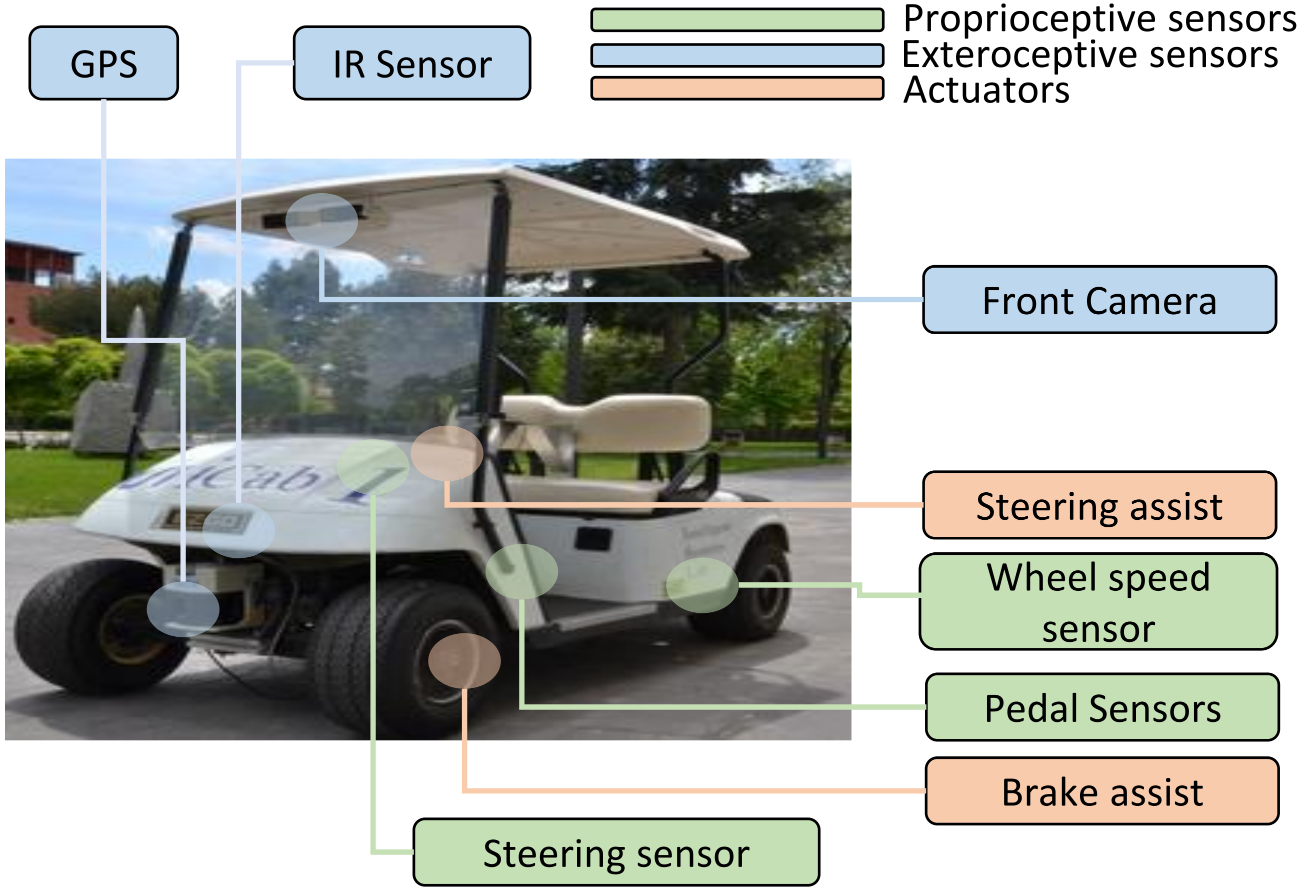
Types of sensors:

- Exteroceptive sensors (e.g., camera)
- Interoceptive sensors (e.g., steering sensor)

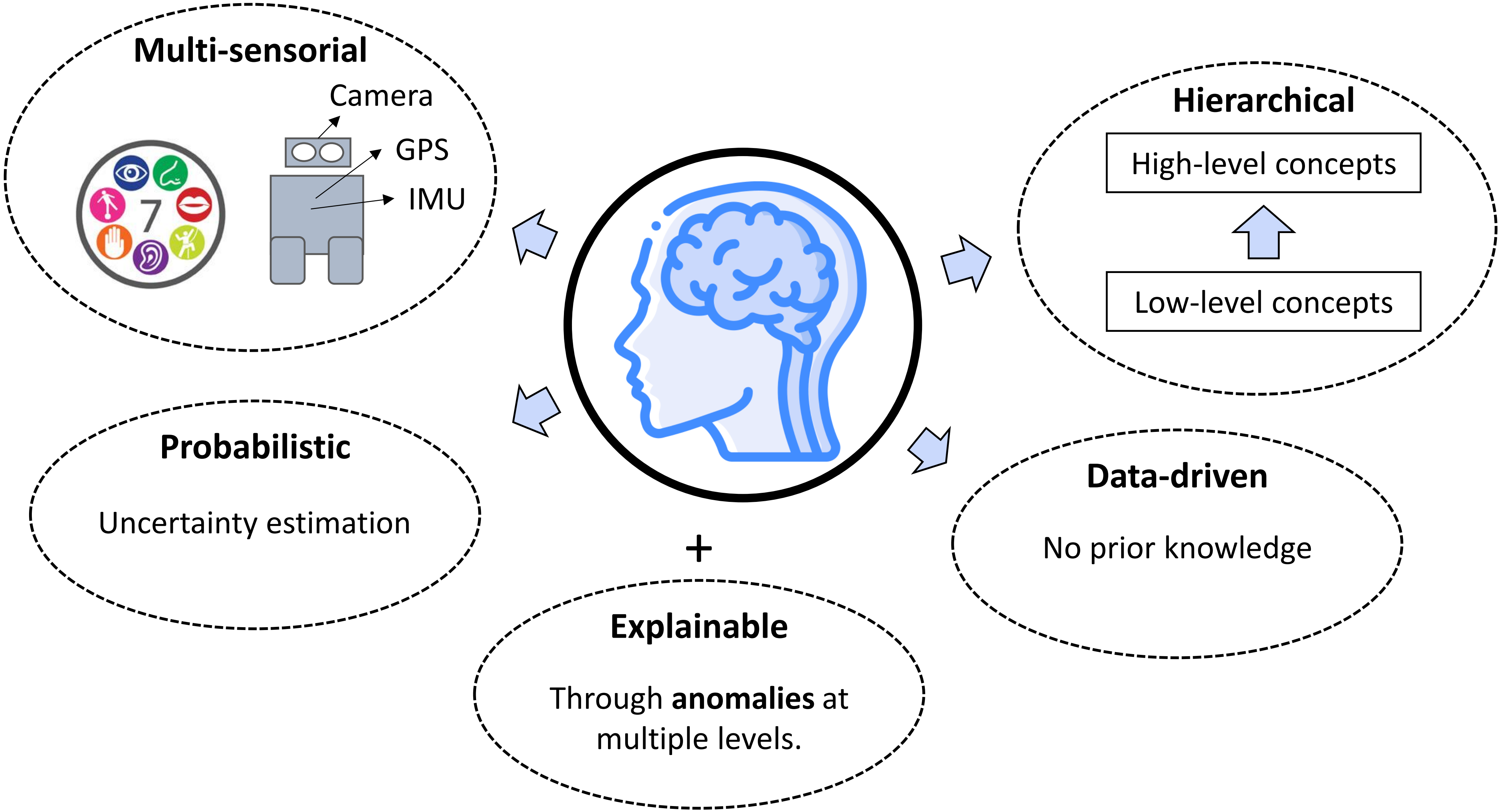
Computationalist vs. **cognitive** approach

Human brain as inspiration

Self-Awareness



Human Reasoning as Inspiration



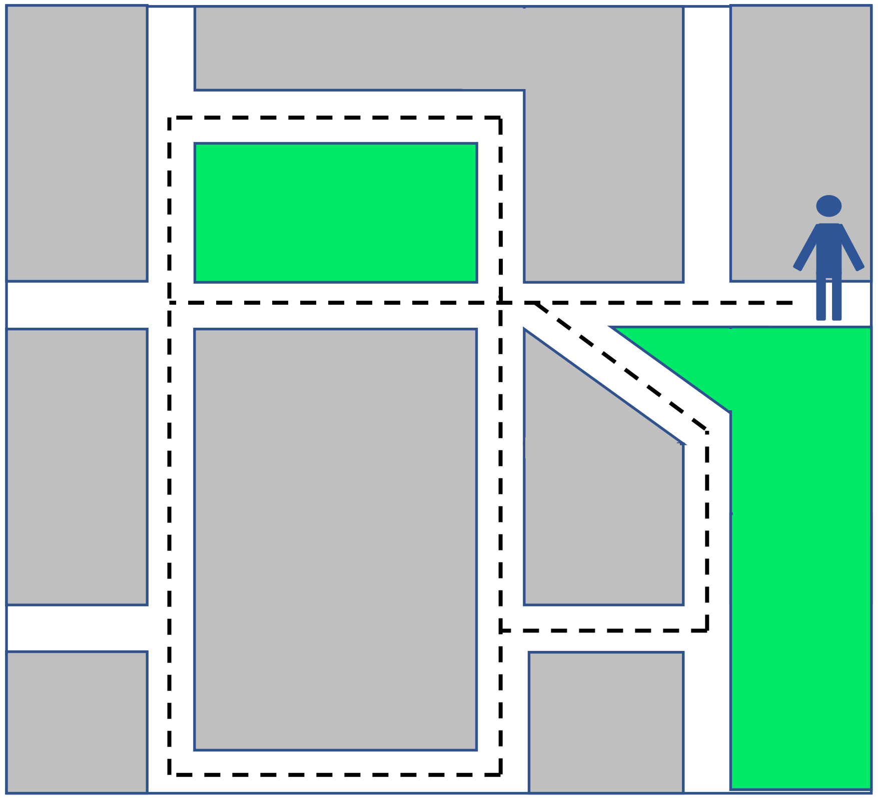
Anomaly Detection

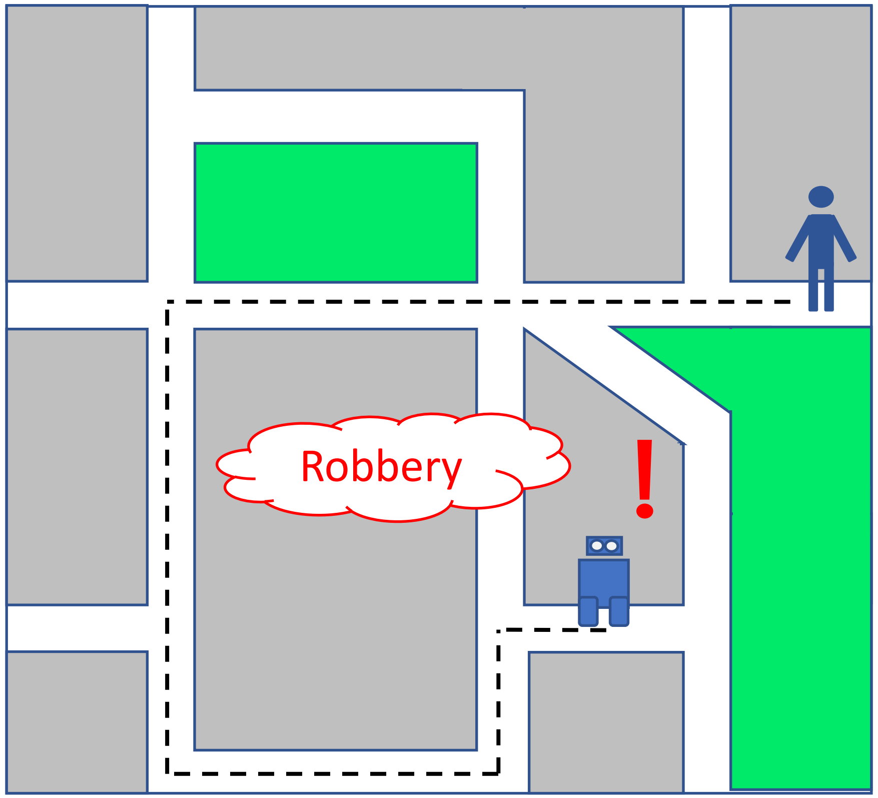
❖ Anomaly detection = process of recognition that an **observation or an experience differs** from observations and experiences learned in the **training phase** of a model.

❖ **Application examples**: video surveillance, medical image analysis, traffic accident detection etc..

[a] V. Mahadevan et al., " Anomaly detection in crowded scenes," CVPR, 2010.
[b] A. Adam et al., "Robust real-time unusual event detection using multiple fixed-location monitors," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 3, 2008
[c] P. Marın-Plaza *et al.*, "Stereo vision-based local occupancy grid map for autonomous navigation in ros," *VISIGRAPP*, 2016.

Anomaly Detection and Localization (1)





Anomaly Detection and Localization (2)

Application examples:

❖ Patrolling robot.

❖ Fault detection.

Images taken from: https://mashable.com/article/singapore-police-patrol-surveillance-robots and https://www.dnv.com/expert-story/maritime-impact/The-drone-squad-for-ship-surveys.html

Comparison with the State of the Art of Cognitive SA Architectures

❖ Few self-awareness approaches have been presented throughout the years, as this area is still in its infancy.

	Probabilistic	Hierarchical	Multi-sensorial	Data-driven	Explainable
[a] (high-level proposal)	✓	✓	✓	✗	✓
[b]	✓	✓	✗	✓	✗
RoboErgoSum [c]	✓	✓	✓	✓	✗
Ours	✓	✓	✓	✓	✓
JEPA [d]	P	✓	✓	✓	✗

JEPA = Joint Embedding Predictive Architecture

P = potentially

[a] L.A. Dennis, M. Fisher, "Verifable self-aware agent-based autonomous systems", Proceedings of the IEEE, vol. 108, n. 7, pp. 1011–1026, 2020.
[b] R. Golombek, S. Wrede, M. Hanheide, M. Heckmann, "Learning a probabilistic self-awareness model for robotic systems", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2745–2750, 2010.
[c] R. Chatila et al., "Toward self-aware robots", Frontiers in Robotics and AI, vol. 5, n. 88, 2018
[d] Y. LeCun, "A Path Towards Autonomous Machine Intelligence", OpenReview Archive, 2022

Theoretical Background

Dynamic Bayesian Networks (DBNs)

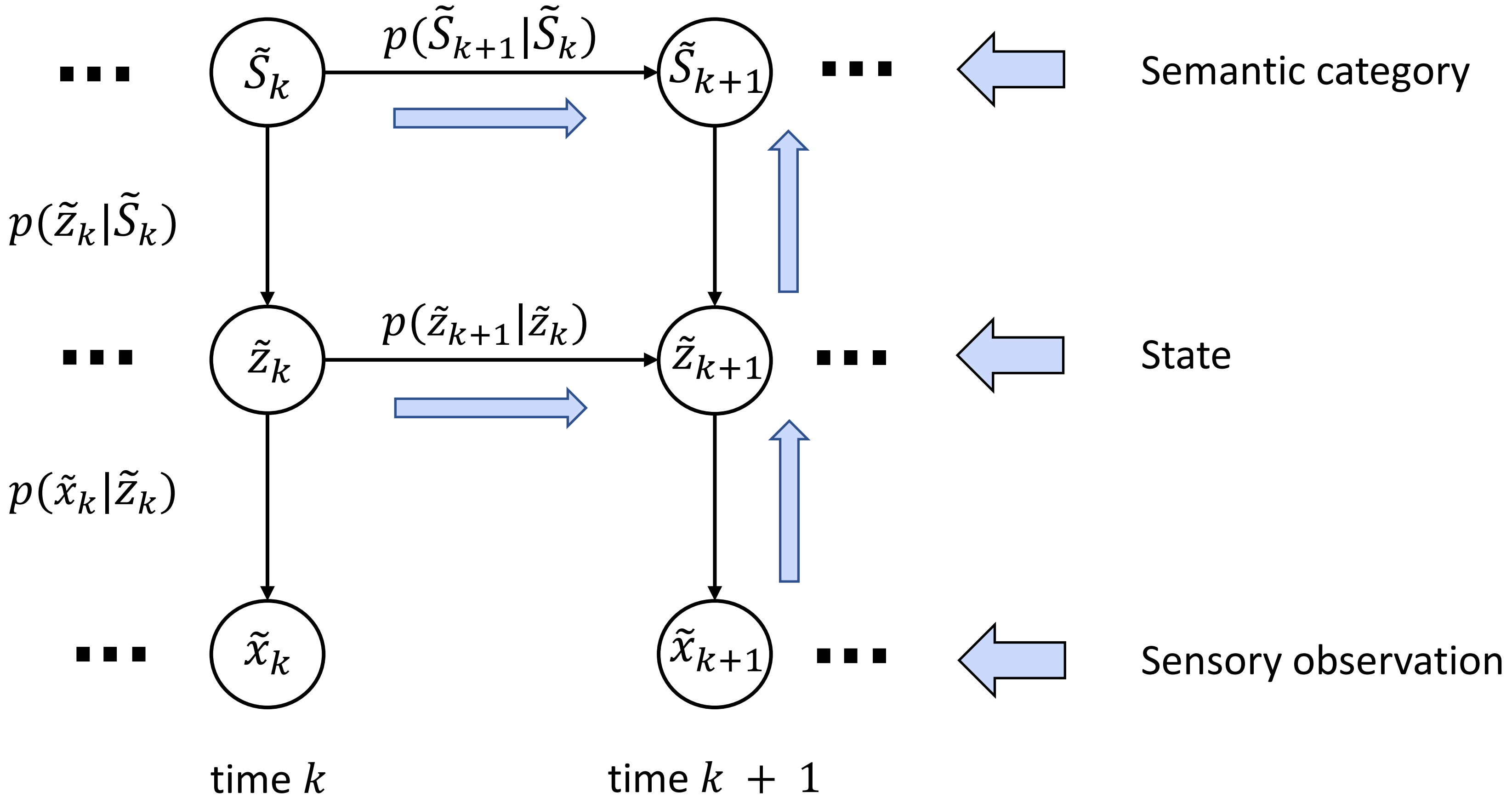
Probabilistic

Hierarchical

Data-driven

Explainable

Multi-sensorial



Learning the Markov Jump Particle Filter

Sensor observations
$\{x_k\}_{k=1\ \dots K}$

Null force filter

$\{\tilde{z}_k\}_{k=1\ \dots K}$

With $z_k = [z_k, \dot{z}_k]$

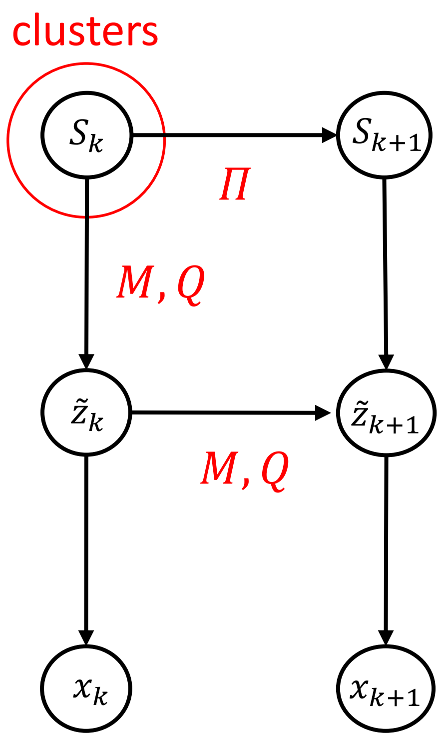
Clustering

MJPF vocabulary:

Cluster mean M;

Cluster covariance Q;

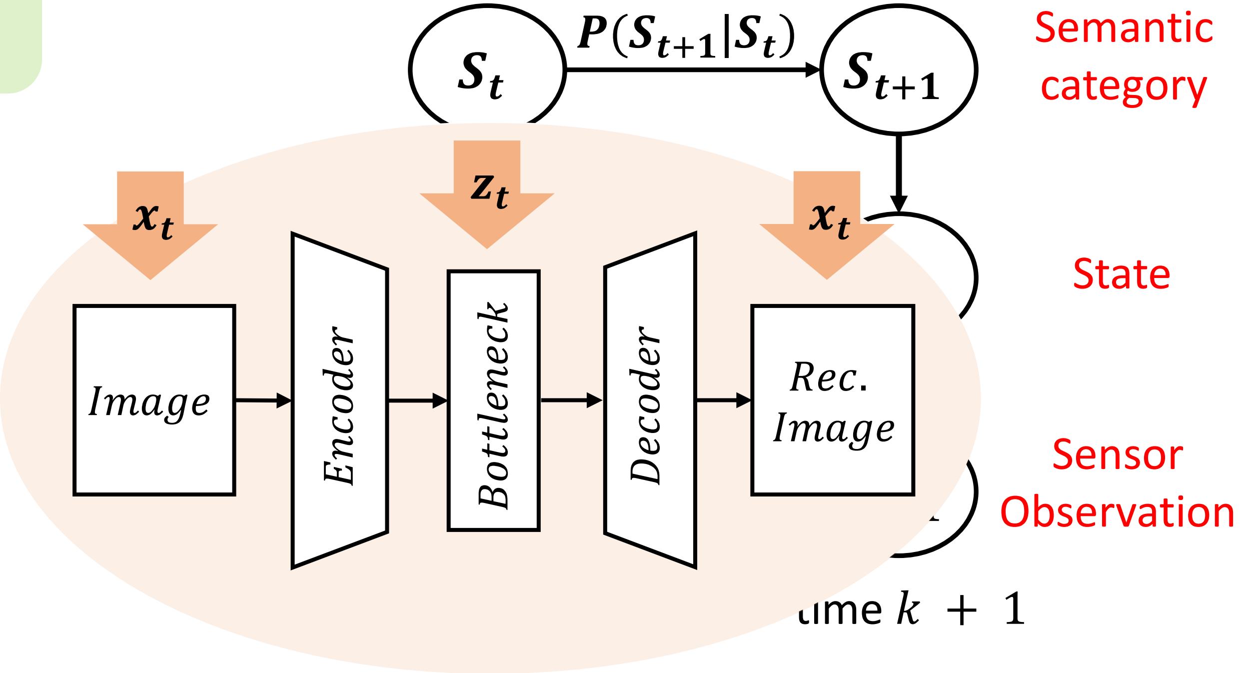
Transition matrix between clusters Π.



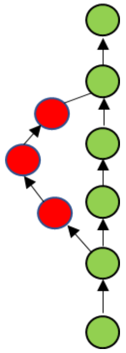
Switching Linear Dynamical Systems for Images

Problem: Switching Linear Dynamical Systems can not be directly applied to data coming from high-dimensional sensors.

Solution: Dimensionality reduction through **Variational Autoencoders**.

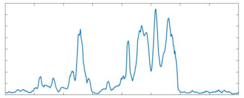


General Architecture [a]



Semantic Anomaly (Kullback Leibler Divergence):

$$D_{KL}(\pi(\tilde{S}_t)||\lambda(\tilde{S}_t)) + D_{KL}(\lambda(\tilde{S}_t)||\pi(\tilde{S}_t))$$



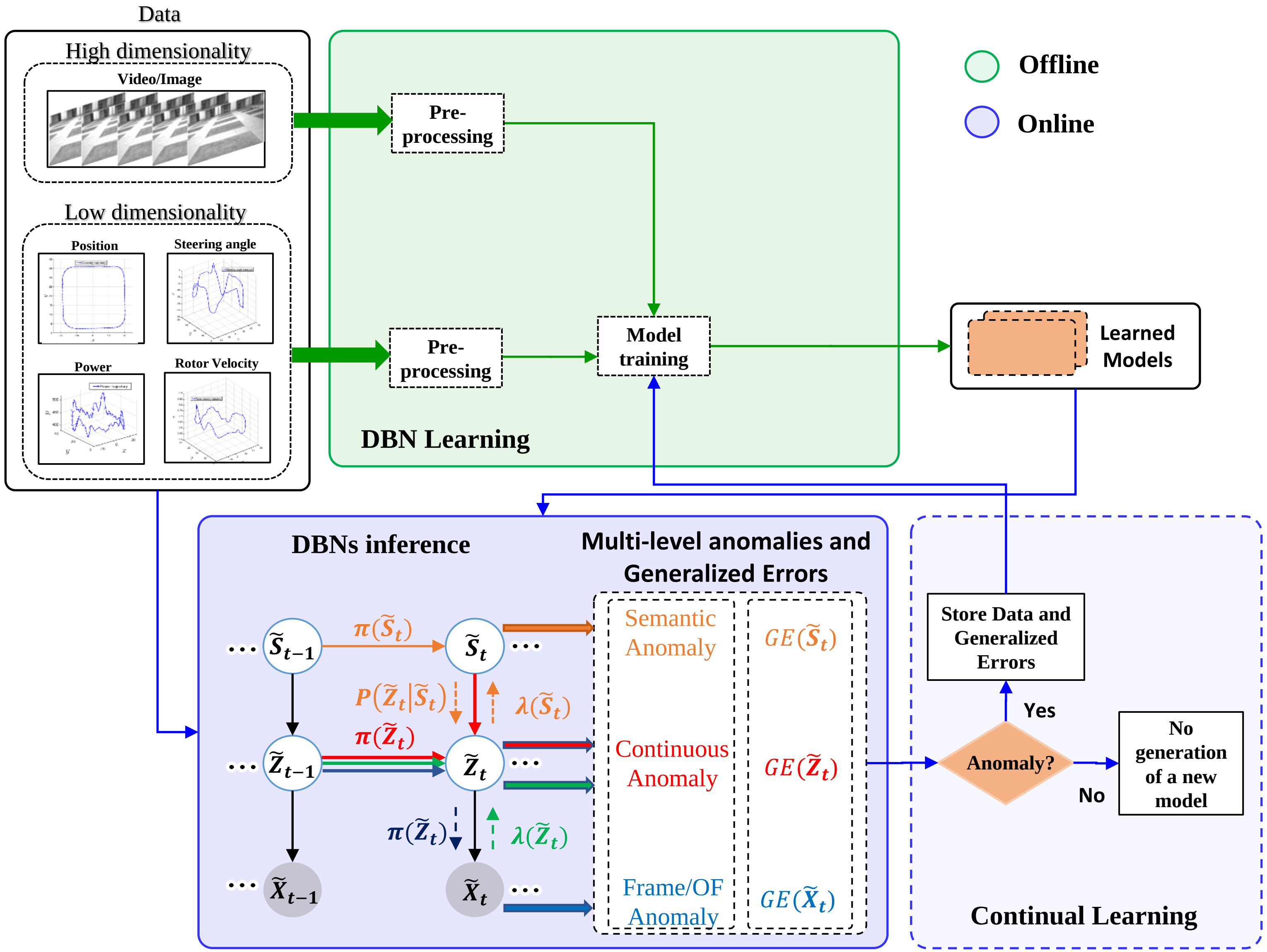
Continuous Anomaly (Bhattacharya):

$$D_B(\pi(\tilde{z}_t), \lambda(\tilde{z}_t))$$

Frame Anomaly (MSE):

$Pred.Error = |TrueImage - PredictedImage|$

$Rec.Error = |TrueImage - ReconstructedImage|$



[a] C. Regazzoni et al., "Probabilistic Anomaly Detection Methods Using Learned Models from Time-Series Data for Multimedia Self-Aware Systems," in "Advanced Methods and Deep Learning in Computer Vision", E. R. Davies, O. Camps, M. Turk, 1st October 2021.

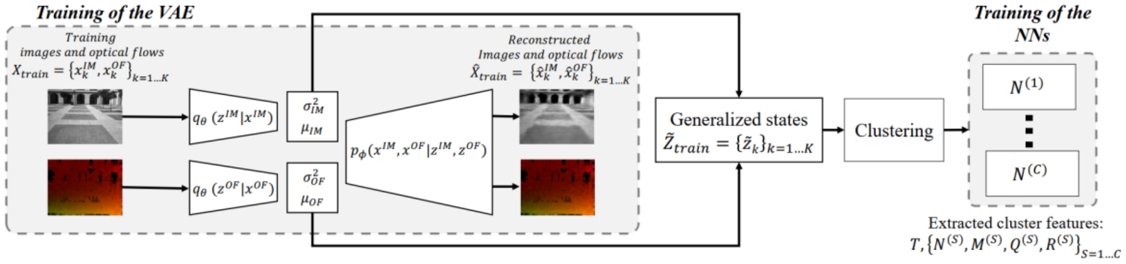
A selection of methods and results

Developed SA Methods

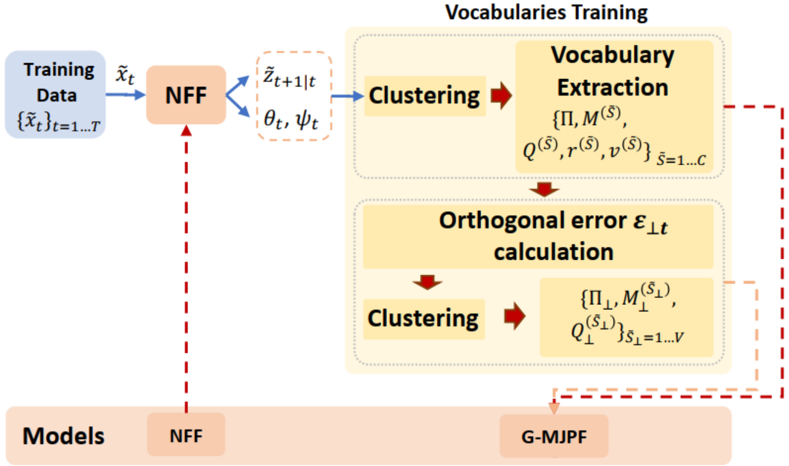
DBC = Driver Behavior Classification

Single-sensor architectures

*Anomaly detection on **video** data*

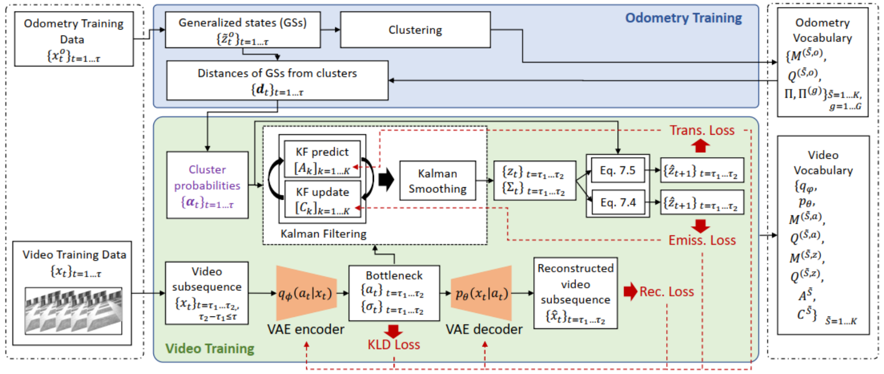


*Anomaly detection on **odometry** data + DBC*

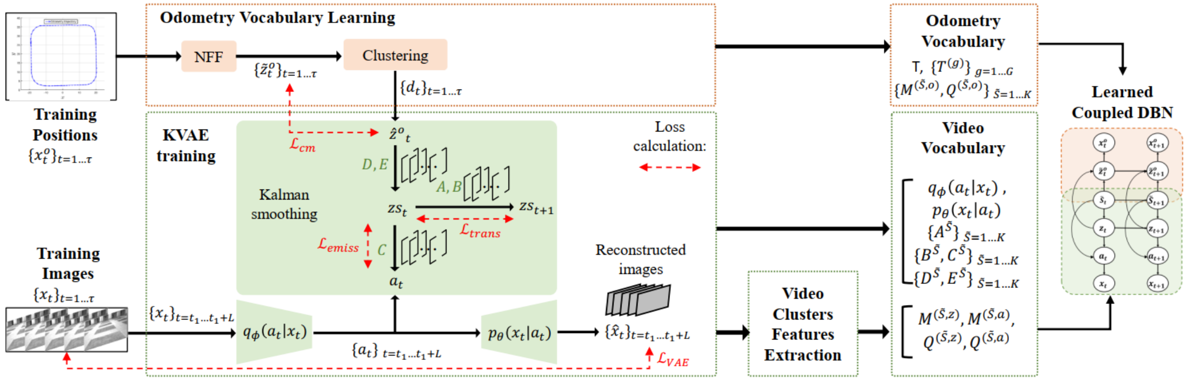


Multi-sensor architectures

*Anomaly detection on **odometry and video** data*



*Anomaly detection on odometry and video data + **localization***



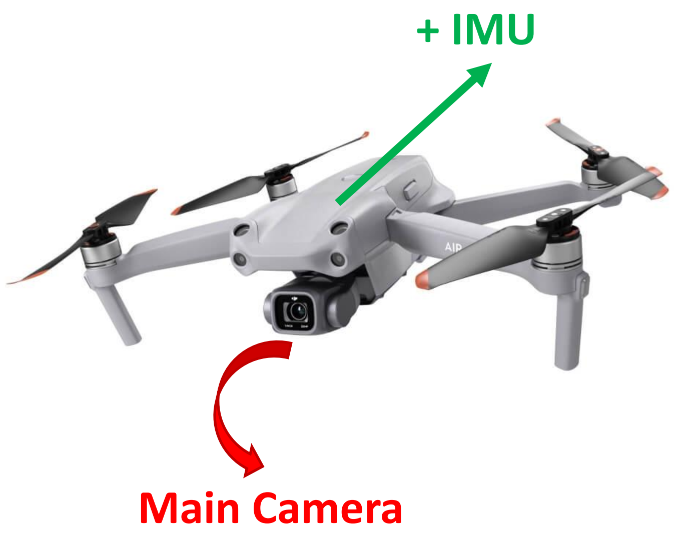
Applicable Data

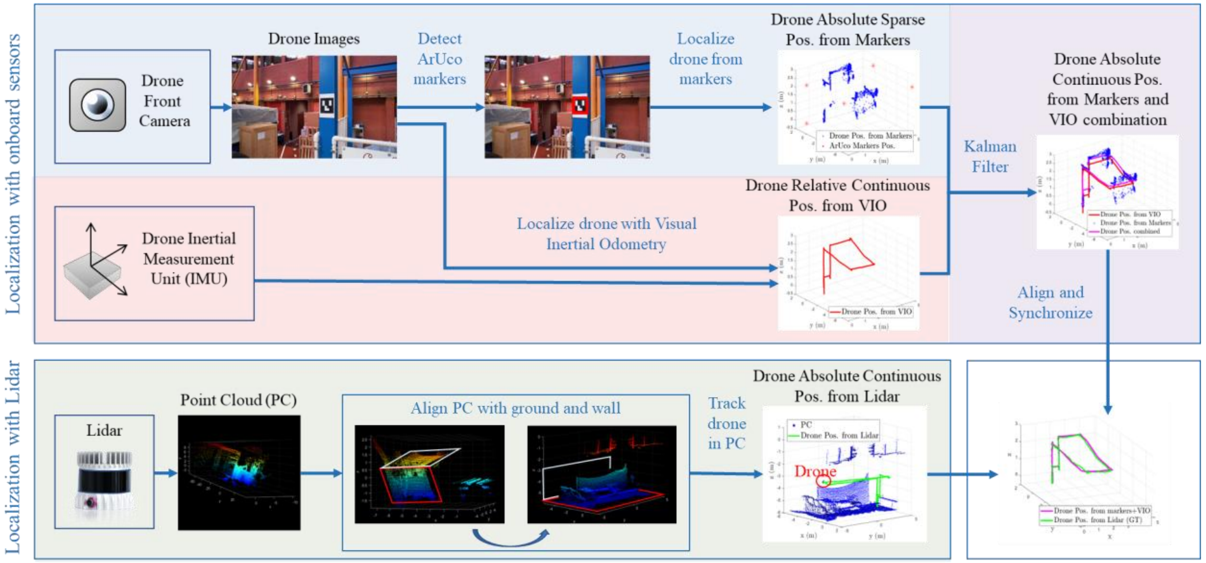
Terrestrial Datasets

Onboard cameras + GPS/IMU:
e. g., iCab, UAH, Carla, Egocart

Fixed cameras:
e.g., Avenue, Subway

Aerial Datasets





Multilevel anomaly detection Through Variational Autoencoders and Bayesian Models for self-aware Embodied Agents

Method Introduction

Objective:

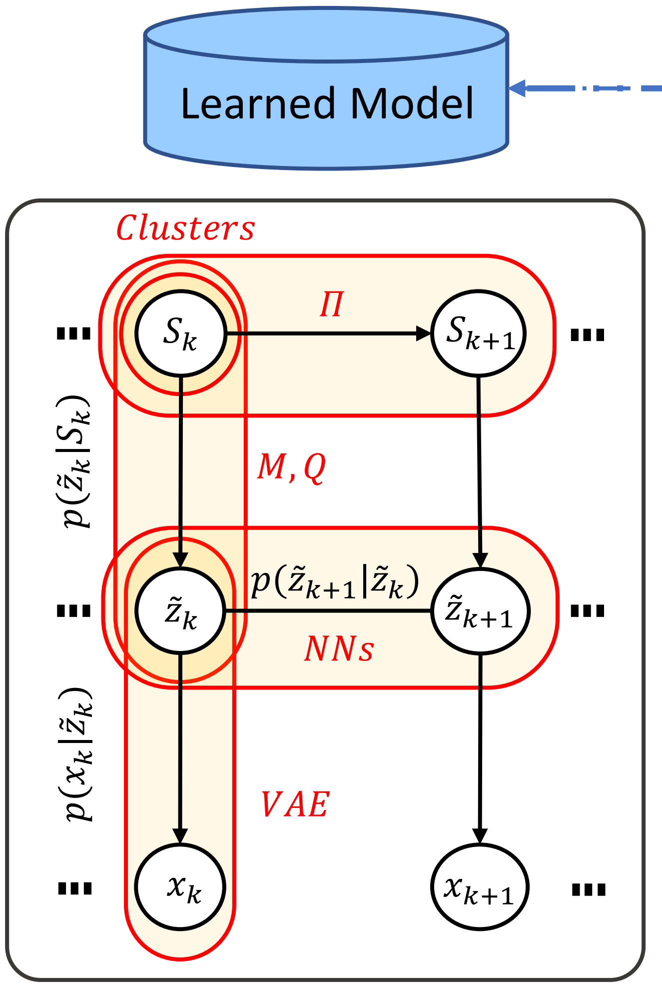
Multi-level anomaly detection performed on **video** data (from static or moving cameras).

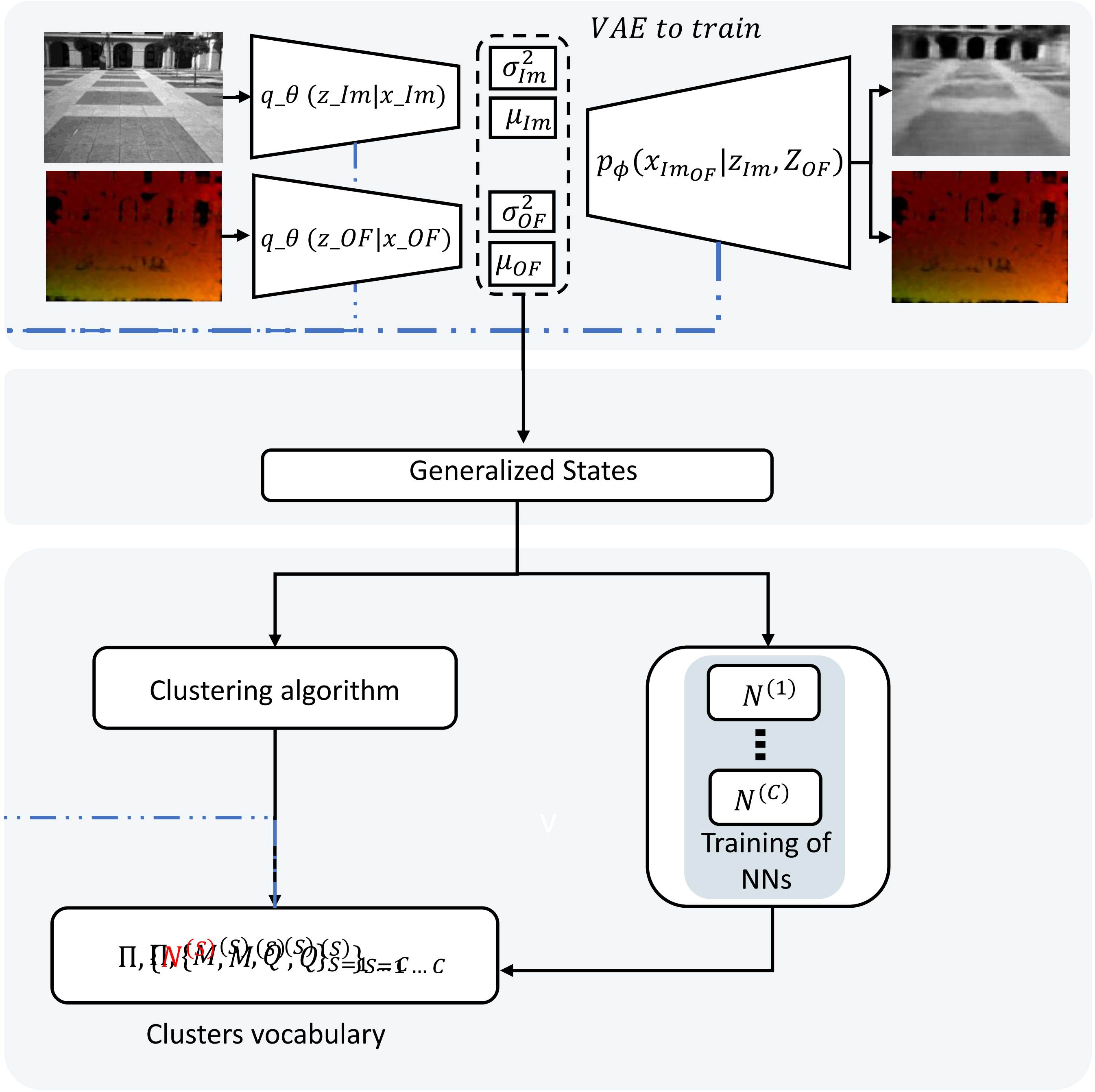
Probabilistic, Data-Driven, Hierarchical, Explainable

Multi-sensorial

Homogeneity with the low-dimensional case

Training Phase





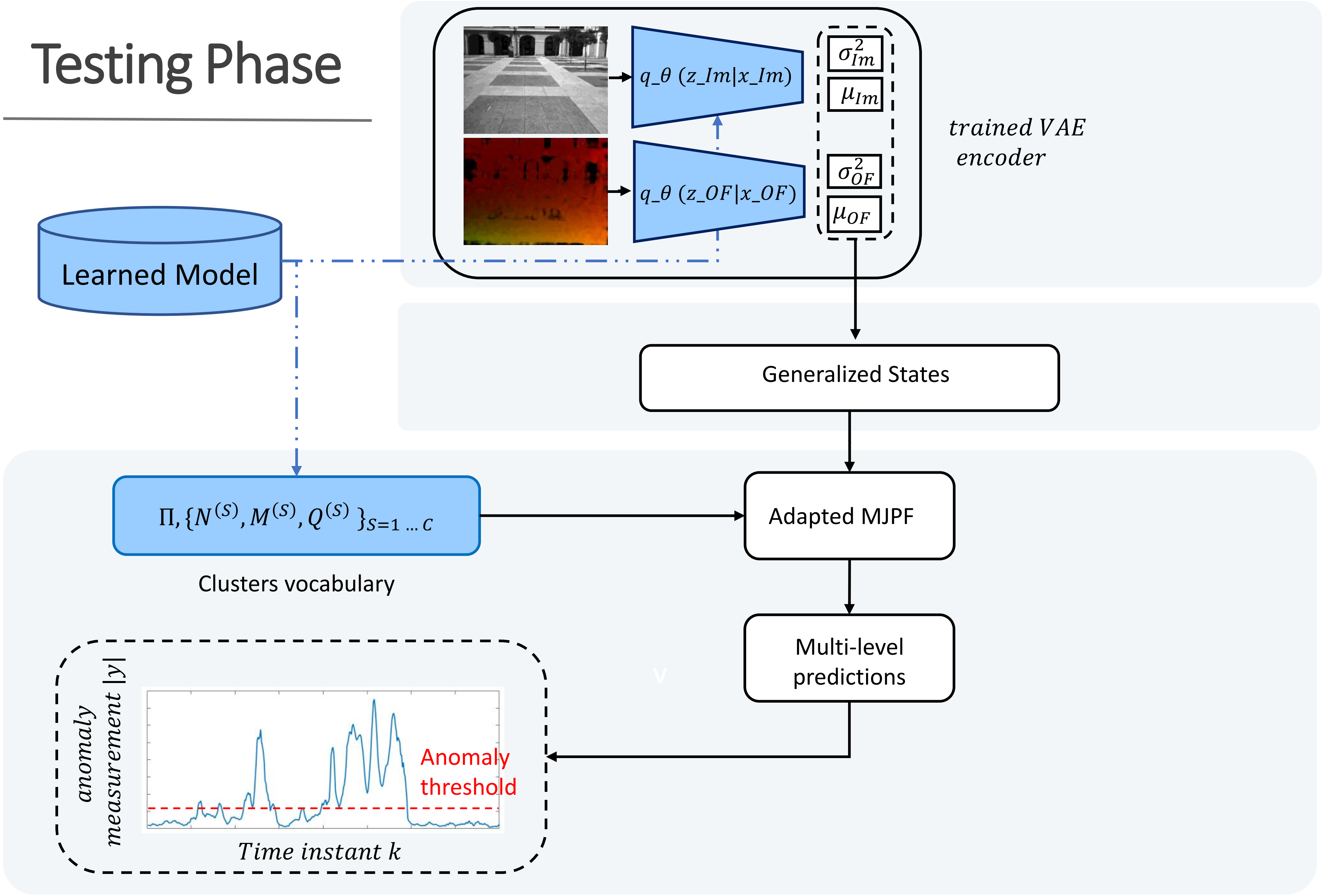
A VAE is trained to reconstruct a set of images.

The Generalized States $\tilde{Z}_{train}$ are obtained.

Clustering is performed, and the feature variables V of each cluster are extracted.

For each cluster, a Neural Network is trained to be later used as a prediction model.

Testing Phase

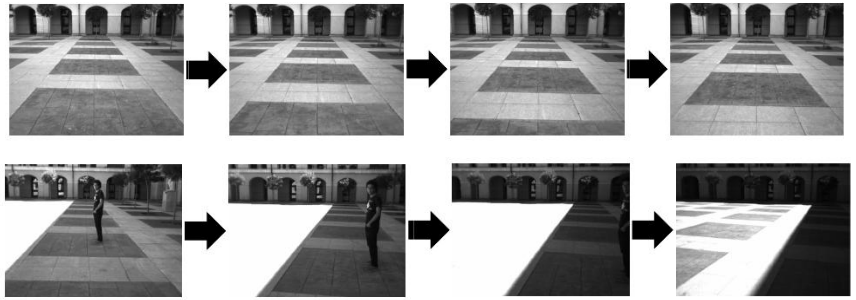


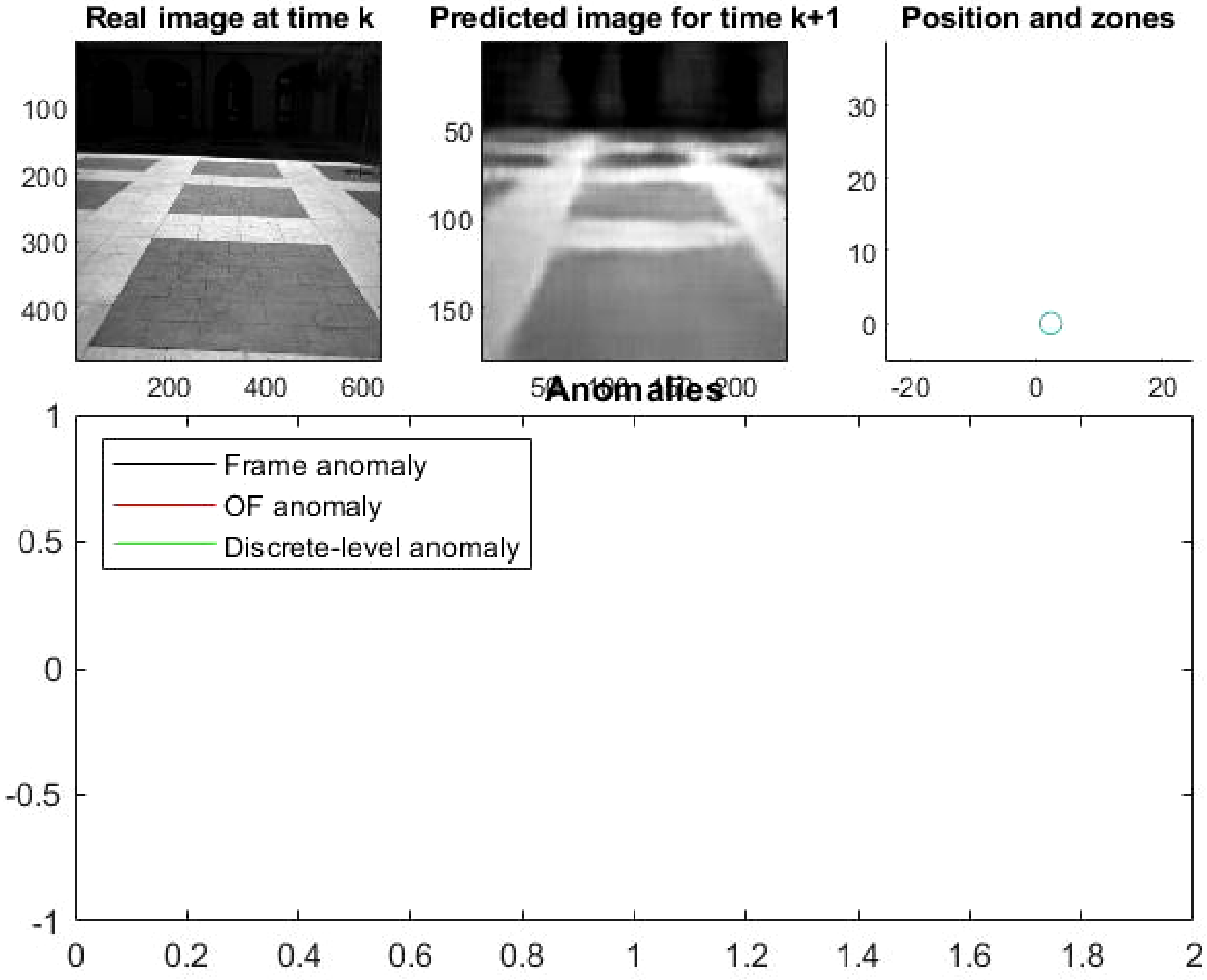
The testing images X_{test} are propagated through the encoder q_θ of the VAE, and a set of latent features vectors μ_{test} and σ^2_{test} are obtained.

The Generalized States $\tilde{Z}_{test}$ are obtained.

An Adapted version of the MJPF is used to make predictions of the next frames and detecting abnormalities at the state level.

Qualitative Results: Anomaly Detection





Quantitative results on various datasets

	GT	AUC Img. Rec. Err.	AUC Img. Pred. Err.	AUC KLDA
Exit	[a], original	0.882	0.896	0.775
	[a], additional	0.865	0.879	0.818
	[b]	0.902	0.910	0.818
Entrance	[b]	0.731	0.732	0.604
	[c]	0.727	0.737	0.626
Avenue	[d]	0.862	0.851	0.671
iCab PA	[a]	0.81	0.87	0.77
iCab U-turn	[a]	0.94	0.91	0.8
iCab ES	[a]	0.82	0.81	0.81

[a] G. Slavic, M. Baydoun, D. Campo, L. Marcenaro, and C. Regazzoni, "Multilevel Anomaly Detection Through Variational Autoencoders and Bayesian Models for Self-Aware Embodied Agents," IEEE Transactions on Multimedia, vol. 24, pp. 1399-1414, 2021

[b] J. Kim, and K. Grauman, "Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates," Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2928, 2009

[c] V. D. de Gevigney, P. Marteau, A. Delhay, and D. Lolive, "Video latent code interpolation for anomalous behavior detection," International Conference on Systems, Man, and Cybernetics (SMC), pp. 3037-3044, 2020

[d] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," IEEE International Conference on Computer Vision (ICCV), pp. 2720-2727, 2013

Comparison with other state-of-the-art methods

Method	Avenue	Exit	Entrance	Year	Interpretability /Explainability	No additional supervision
[a]	0.702	0.807	0.943	2016	✗	✓
[b]	0.803	0.940	0.847	2017	✗	✓
[c]	0.892	0.946	0.902	2019	✗	✓
[d]	0.823	0.932	0.806	2020	✗	✓
Ours	**0.862**	**0.910**	**0.732**	**2021**	**✓**	**✓**
[e]	0.866	-	-	2021	✓	✗
[f]	0.883	-	-	2022	✓	✗
[g]	0.860	-	-	2023	✓	✗

[a] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences", IEEE Conference on Computer Vision and Pattern Recognition, pages 733–742, 2016

[b] Y. S. Chong, and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder", Advances in Neural Networks - International Symposium on Neural Networks, vol. 10262, pages 189–196, 2017

[c] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos ", IEEE Transactions on Multimedia, vol. 22, n. 8, pp. 2138–2148, 2020

[d] V. D. de Gevigney, P. Marteau, A. Delhay, and D. Lolive, "Video latent code interpolation for anomalous behavior detection," International Conference on Systems, Man, and Cybernetics (SMC), pp. 3037-3044, 2020

[e] S. Szymanowicz, J. Charles, and R. Cipolla, "X-MAN: explaining multiple sources of anomalies in video", In IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 3224–3232, 2021

[f] S. Szymanowicz, J. Charles, and R. Cipolla, "Discrete neural representations for explainable anomaly detection", In IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1506–1514, 2022

[g] A. Singh, M. J. Jones, and E. G. Learned-Miller, "EVAL: explainable video anomaly localization", In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18717–18726, 2023

Vehicle Localization and Anomaly Detection for Video Surveillance in a Dynamic Bayesian Network Framework

Method Introduction

Objectives:

Multi-level anomaly detection performed on **video and odometry** data.

+ Visual-Based Localization.

Probabilistic, Data-Driven, Hierarchical, Explainable**, Multi-sensorial**

Increased homogeneity with the low-dimensional case

Training Overview

Odometry data $\{x_t^o\}_{t=1\ldots\tau}$

Null-Force Filter

$\{\tilde{z}_t^o\}_{t=1\ldots\tau}$, $\tilde{z}_t^o = [z_t^o, \dot{z}_t^o]$

Camera data $\{x_t\}_{t=1\ldots\tau}$

KVAE training

VAE training

Images latent states $\{a_t\}_{t=1\ldots\tau}$

Learning of KVAE matrices

Vocabulary Learning: clustering of $\{\tilde{z}_t^o, a_t\}_{t=1\ldots\tau}$

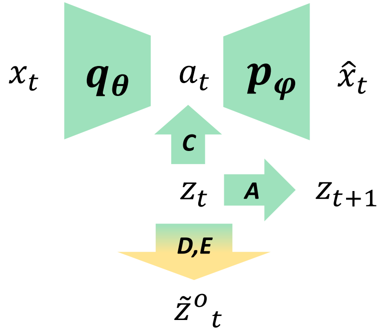
Odometry Vocabulary (K clusters)

- Cluster means $M^{(\tilde{S}_t,o)}$
- Cluster covariances $Q^{(\tilde{S}_t,o)}$
- Transition matrix Π
- Temporal transition matrices $T^{(g)}, g = 1 \ldots G$

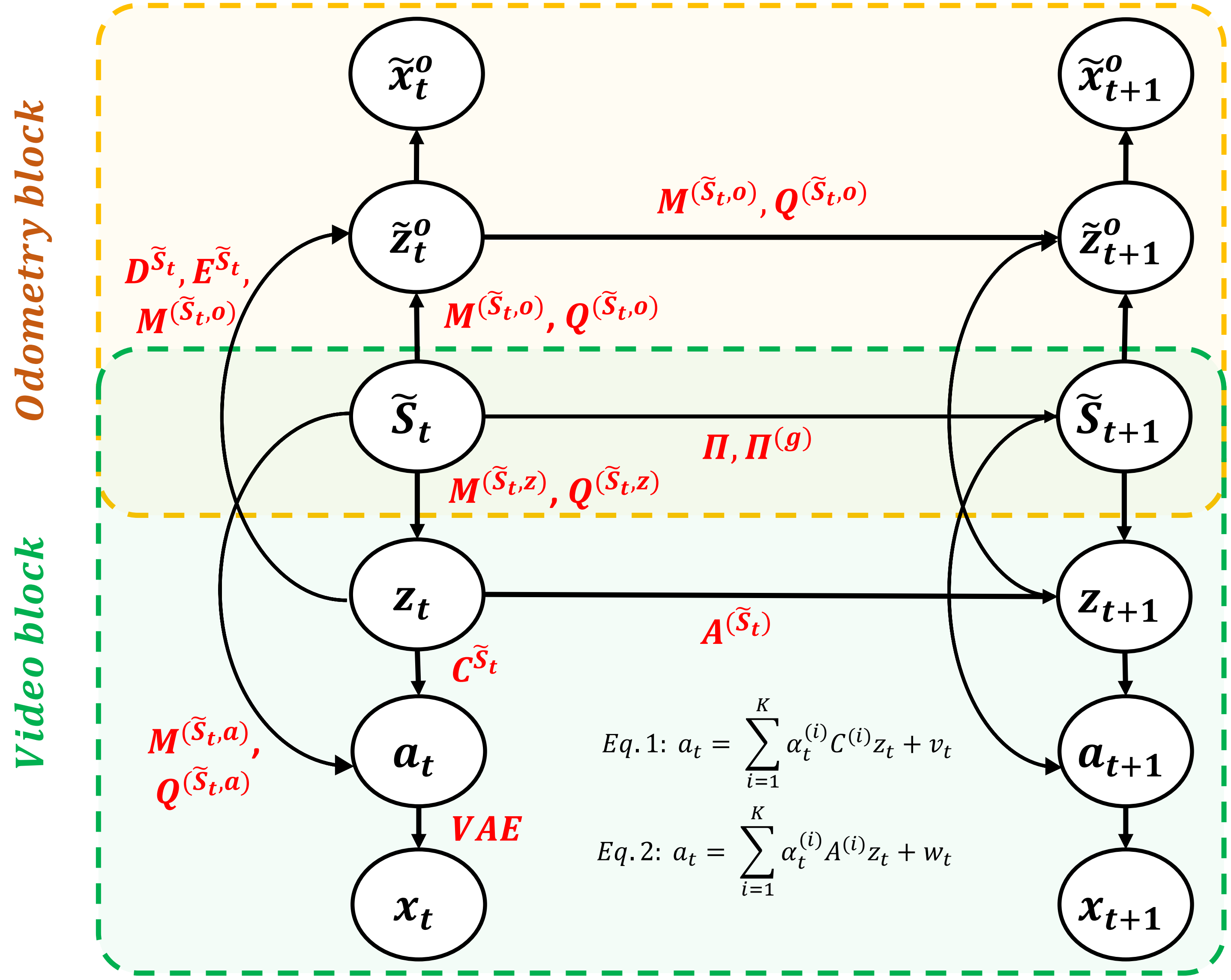
$\left\{d_t^{(k)}\right\}_{k=1\ldots K,\, t=1\ldots T}$

Video Vocabulary

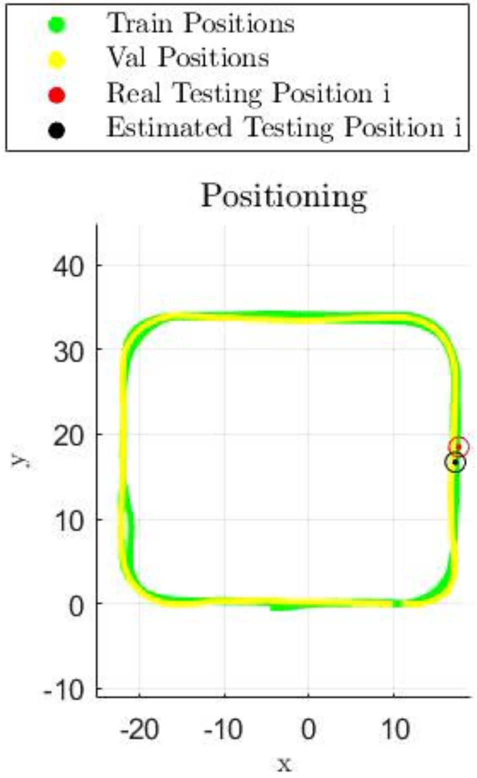
- VAE's encoder q_θ and decoder p_φ
- Cluster means $\left(M^{(\tilde{S}_t,z)}, M^{(\tilde{S}_t,a)}\right)$
- Cluster covariances $\left(Q^{(\tilde{S}_t,z)}, Q^{(\tilde{S}_t,a)}\right)$
- Emission matrices $\left\{C^{(k)}\right\}_{k=1\ldots K}$
- Transition matrices $\left\{A^{(k)}\right\}_{k=1\ldots K}$
- Cross-modality matrices $\left\{D^{(k)}, E^{(k)}\right\}_{k=1\ldots K}$

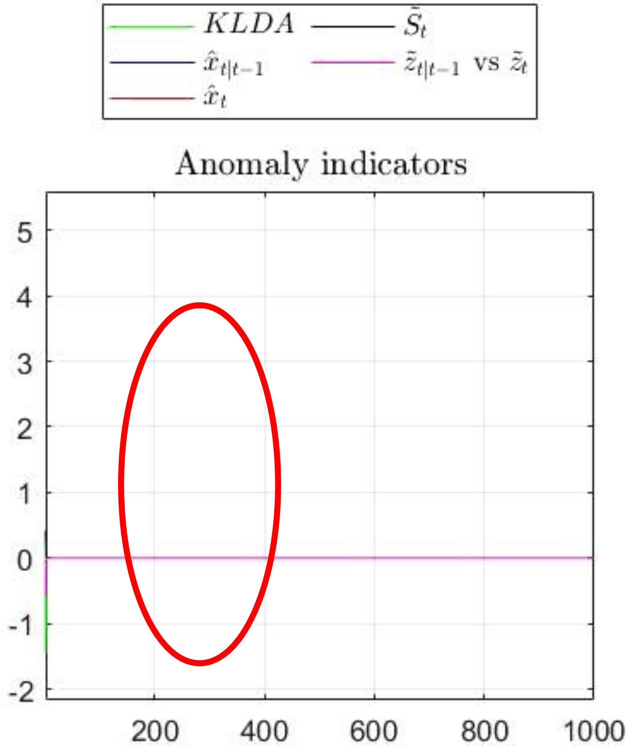


Coupled Dynamic Bayesian Network



Positioning and Anomaly Estimation Example





Quantitative results and comparisons

		Egocart		iCab Emergency Stop		Drone Frontal Motion		Drone Lateral Motion	
		Mean Err (m)	Median Err (m)	Mean Err (m)	Median Err (m)	Mean Err (m)	Median Err (m)	Mean Err (m)	Median Err (m)
Methods without pre-trained models	**IR-VAE**	1.60	0.32	23.00	23.00	0.20	0.16	0.47	0.25
	IR-TC-VAE	3.61	0.39	23.00	23.00	0.20	0.16	0.86	0.33
	REG-ENC	8.59	7.66	23.88	22.78	0.89	0.76	1.83	1.14
	Ours	**1.65**	**0.96**	**0.98**	**0.75**	**0.23**	**0.14**	**0.87**	**0.38**
Methods with pre-trained models	**IR-IV3 [a]**	0.73	0.28	1.28	0.61	0.18	0.16	0.32	0.20
	IR-TC-IV3 [a]	-	-	0.72	0.60	0.18	0.16	0.32	0.20
	IR-PNET-VGG16 [a]	2.17	1.38	-	-	-	-	-	-
	IR-TC-VGG16 [a]	0.52	0.28	-	-	-	-	-	-
	IR-TR-TC-VGG16 [a]	0.44	0.29	-	-	-	-	-	-
	REG-SVR-PNET-RGB-VGG16 [a]	1.96	1.54	-	-	-	-	-	-
	REG-PNET-RGB-POS-IV3 [a]	0.42	0.29	1.52	1.15	0.24	0.20	0.74	0.71

IR = image retrieval; TC= Temporal Constraint; ENC = encoder; PNET = PoseNet; SVR = Support Vector Regression; POS = position

[a] E. Spera, A. Furnari, S. Battiato, and G. M. Farinella, "Egocart: a benchmark dataset for large-scale indoor image-based localization in retail stores," IEEE Trans. Circuits Syst. Video Technol., vol. 31, pp. 1253–1267, Sept. 2021.

Conclusions and Future Work

Conclusions

- The development of self-awareness architectures for autonomous vehicles inspired from **human reasoning**, and that incorporate characteristics such as being **probabilistic, hierarchical, data-driven, explainable, and multi-sensorial**;

- The use of **anomaly detection** inside this architecture to identify **new rules** that continually emerge from the data and that indicate the necessity to build a **new model;**

- The employment of **low and high dimensional data**, which should be handled as **homogeneously** as possible;

- The **localization** of the vehicle in the environment, as an additional capability of the architecture .

Future Work

- ❖ Closing the Continual Learning cycle;
- ❖ Further explaining the anomalies;
- ❖ Further analyzing the anomalies;
- ❖ Inserting other sensory modalities;

Thank you for your attention