# A MULTI-PERSPECTIVE APPROACH TO ANOMALY DETECTION FOR SELF-AWARE EMBODIED AGENTS

*Mohamad Baydoun*[1,2]*,Mahdyar Ravanbakhsh*[1]*,Damian Campo*[1]*,Pablo Marin*[3]*, David Martin*[3]*, Lucio Marcenaro*[1]*, Andrea Cavallaro*[2]*, Carlo S. Regazzoni*[1,3]

[1]DITEN, University of Genoa, Italy. [2]CIS, Queen Mary University of London, UK.
[3]Carlos III University of Madrid, Spain.

## ABSTRACT

This paper focuses on multi-sensor anomaly detection for moving cognitive agents using both external and private first-person visual observations. Both observation types are used to characterize agents motion in a given environment. The proposed method generates locally uniform motion models by dividing a Gaussian process that approximates agents displacements on the scene and provides a Shared Level (SL) self-awareness based on Environment Centered (EC) models. Such models are then used to train in a semi-unsupervised way a set of Generative Adversarial Networks (GANs) that produce an estimation of external and internal parameters of moving agents. Obtained results exemplify the feasibility of using multi-perspective data for predicting and analyzing trajectory information.

***Index Terms—*** Abnormality detection, Gaussian process, Generative adversarial networks , Situational awareness, multi-sensor systems

## 1. INTRODUCTION

Fully autonomous systems need perception to navigate through scenes and recognize objects in real environments [1]. Recent advances in signal processing and machine learning techniques can be useful to design autonomous systems equipped with a self-awareness module that facilitates to recognize contextual information while a given task is executed. The capability of detecting abnormal situations based on such self-awareness is an important task that allows autonomous systems to increase their situational awareness and the effectiveness of the decision making submodules [2].

The analysis of observed moving agents for understanding normal/abnormal dynamics in a given scene represents an emerging research field [3, 2, 4]. This paper proposes a methodology for abnormality detection based on multiple sensors that observe the same phenomenon from different perspectives. Abnormalities can be first detected as deviations from Environment Centered (EC) models, i.e., from an observer viewpoint which does not have access to internal agent variables. Such layer can be defined as a Shared Level (SL) of self-awareness, since the observed information, e.g., observed position and velocity can be measured easily from an external observer.

An observed agent can also have further information corresponding to what it can observe from a first person viewpoint (FPV) while a task is performed. Abnormalities related to unexpected observations acquired while performing a task can be considered as the essential information to define a Private Layer (PL) of self-awareness: Such experiences are available only to the agent itself. Accordingly, an external observer cannot access to such information and has to rely solely on SL information.

Analyzing phenomena from different sensory data is definitively not a new problem. In [5], researchers use video data together with orientation information to capture the 3D motion of a human body. In [6], a multi-sensor monitoring system is proposed to prevent accidents and detect falls. Additionally, several researchers used multiple-cameras to recognize abnormalities [7, 8].

One of the main novelties of this work consists of a strategy which processes data from first and external viewpoints and facilitates a subdivision of subject's behaviors into basic dynamical models (activities). Dynamic normality models and related algorithms can detect abnormalities by fusing shared and private agent information. Identified anomalies are here defined as patterns that have not seen or learned in previous experiences [9, 3] by taking into consideration private and shared perspectives of the same phenomenon.

Observations acquired from the external viewpoint (EV) are composed of agents' positions and velocity with respect a fixed reference system. Locations and velocities (actions) are analyzed by a Gaussian Process (GP) regression that estimates the most probable agent's action in each position of the whole scene. Results from the GP regression are then clustered in zones through a superpixel algorithm approach introduced in [10] that encodes action patterns, e.g., going straight or curving.

Images collected synchronously from an FPV are used to compute optical flow at each frame. Video sequences can be seen as PL data related to the reference system of the moving agent itself. The optical flow between video frames can be seen as a private representation of the agent's action in the PL. Accordingly, in order to understand the relation between a given FPV image and its optical flow, a Generative Adversarial Nets (GANs) approach [11] is adopted for training deep networks in a supervised way. Supervision here consists in supposing that the action patterns previously obtained by the GP approach for defining EC normality can be used to train normal visual models.

Both SL and PL learned models can be used to predict the dynamics of a vehicle performing a task. Produced models by each approach can be used to describe how these two representations are related to each other in normal conditions. This paper shows that the PL representation learned in a supervised way can provide further normality descriptors, enriching the ones obtained in an unsupervised way from GP for EC normality representation.

The remainder of the papers is organized as follows: section 2 describes the GP approach (section 2.1) and the GAN (section 2.2.1) for SL and PL self awareness, respectively. Section 3 shows the dataset that was used for obtaining results that are detailed in sections 4.1 and 4.2. Conclusions and future research directions are presented in section 5.

## 2. PROPOSED METHOD

### 2.1. Representation of observed dynamic motion

To model the SL self-awareness, we use a state space representation from an external observer placed in the EC reference system. Accordingly, the state subspace $X$ represents the location of an agent in the environment whereas $\dot{X}$ represents its velocity. the whole scene can be seen as a grid of possible locations where agents can be. Let such spatial grid be defined as $\tilde{X} = \{X^1, X^2, ..., \dot{X}^M\}$. As can be seen, $\tilde{X}$ is a set of $M$ locations that cover the whole environment. Given a set of observations from a moving agent, it is possible to express its positions and velocities as the subsets $X^* = \{X_1, X_2, ..., \dot{X}_K\}$ and $\dot{X}^* = \{X_1, X_2, ..., \dot{X}_K\}$, where $K$ is the total number of observations of the agent. By using $X^*$ and $\dot{X}^*$, it is possible to use a GP approach [9] that approximates velocity information over spatial grid, $\tilde{X}$, such that:

$$\tilde{\dot{X}} = g(\tilde{X}) + v, \tag{1}$$

where $\tilde{\dot{X}}$ represents an estimation of velocity information for each point of the spatial grid, $g(\cdot)$ takes location information and estimates the expected motion (action) at such position for a given activity. Since agents' actions can be seen as velocities, it is possible to describe them as a GP, where $v$ is a Gaussian zero-mean white noise.

In this work, a 2-dimensional case is considered. Therefore, spatial coordinates and time derivatives consist of two components each, $(x, y)$ and $(\dot{x}, \dot{y})$, respectively. Accordingly, it is possible to represent $(x, y)$ as the pixels of an image whose corresponding color values carry information about agents' actions $(\dot{x}, \dot{y})$. In particular, here RGB images are considered, where Red and Blue colors encode $(\dot{x}, \dot{y})$ respectively and the Green channel is disregarded.

Uncertainties generated by the GP are used to remove information where not enough evidence is observed such as explained in [10]. Since an image that encodes the GP is available, a superpixel algorithm [12] is applied to discretize the image plane into $N$ zones. Each of these zones is characterized by a quasi-constant velocity model $[\dot{x}_n, \dot{y}_n]$, where $n$ indexes identified zones, i.e., $n \in \{1, 2, \ldots, N\}$. Finally, a linear dynamic model can be defined for each zone such as follows:

$$A_n := X_{k+1} = X_k + \Delta k U_{n,k} + w_m, \tag{2}$$

where $U_{n,k} = [\dot{x}_n, \dot{y}_n]^T$, $k$ indexes the time, $\Delta k$ is the sampling time and $w_m$ is the process noise. The variable $U_{n,k}$ is a control input that encodes the action (motivation) of the agent. The process for identifying zones where quasilinear models are valid is summarized in the block diagram in Fig.1.

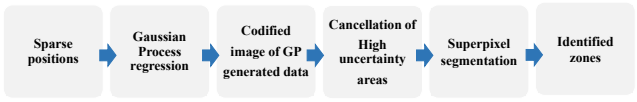



**Fig. 1**: Identification of dynamic motion in terms of zones

#### 2.1.1. Abnormality detection by using Kalman filter method

It is possible to build a set of Kalman Filters (KFs) based on the built dynamical models shown in equation (2). Each KF is designed for tracking linear motions with low error (innovation) when observed data follows already characterized (normal) behaviors inside identified zones.

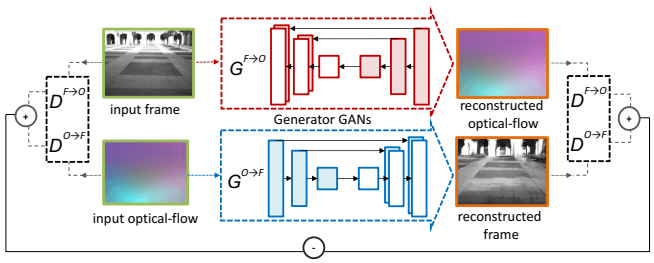



**Fig. 2**: The two GANs structure

As is well known, KFs' innovations represent residual values produced by measurements while assuming a specific normal model. Such values can be used to express abnormalities since they quantify the deviations from normal learned models in the environment. Innovations can be expressed as:

$$\epsilon_{k,n} = Z_k - \hat{X}^n_{k|k-1}, \tag{3}$$

where $\epsilon_{k,n}$ is the innovation generated in the zone $n$ where the agent is located. $Z_k$ represents observed spatial data and $\hat{X}^n_{k|k-1}$ is the KF estimation of the agent's location at the future time $k$ calculated in the time instant $k-1$ (2).

In this work, innovation vectors are composed of two components (one for each axis) and the magnitude of those vectors can be considered as a final measure of abnormality, $\xi$, assuming that the observed agent is inside the region $n$, it is possible write:

$$\xi_k = ||\epsilon_{k,n}||_2,$$

In order to evaluate if an observation is abnormal with respect to the current bank of KFs that encodes the normality in an environment (eq. (2)), an error threshold $\xi_{thres}$ is defined for distinguish between abnormal and normal behaviors at each time instant. Accordingly, if a certain $\xi_k$ exceeds such threshold, the system considers the current measurement $Z_k$ as an anomaly.

### 2.2. Representation of the agent embodied self awareness

In order to represent the PL of agent self-awareness, Generative Adversarial Networks (GANs) [11] are proposed to learn the normality pattern of the observed scene. GANs are deep networks commonly used to generate data (e.g., images) and are trained using only unsupervised data. The supervisory information in a GAN is indirectly provided by an adversarial game between two independent networks: a generator ($G$) and a discriminator ($D$). During training, $G$ generates new data and $D$ tries to understand whether its input is real (i.e., it is a training image) or produced by $G$. The competition between $G$ and $D$ is helpful for boosting the ability of both $G$ and $D$.

#### 2.2.1. Learning the normal pattern of the observed scene

Two channels are used to learn the normal pattern of the observed scene: appearance (i.e., raw-pixels) and motion (optical flow images) for two cross-channel tasks. In the first task, optical-flow images are generated from the original frames. In the second task, appearance information is estimated from an optical flow image.

Let $F_t$ be the $t$-th frame of a training video and $O_t$ the optical flow obtained using $F_t$ and $F_{t+1}$. $O_t$ is computed using [13]. Fig.2 shows two networks: $\mathcal{N}^{F \to O}$, which is trained to generate optical-flow from frames (task 1) and $\mathcal{N}^{O \to F}$, which generates frames from

optical-flow (task 2). In both cases, inspired by [14, 15], our networks are composed of a conditional generator $G$ and a conditional discriminator $D$. $G$ takes as input an image $x$ and a noise vector $z$ (drawn from a noise distribution $\mathcal{Z}$) and outputs an image $r = G(x, z)$ of the same dimensions of $x$ but represented in a different channel.

Both $G$ and $D$ are fully-convolutional networks, composed of convolutional, batch-normalization layers and ReLU nonlinearities. In case of $G$, we adopt the U-Net architecture [14], which is an encoder-decoder. $D$ is proposed to be a *PatchGAN* discriminator [14], which is based on a "small" fully-convolutional discriminator $\hat{D}$. Additional details about the training procedure can be found in [14, 15]. During training, the output of $\hat{D}$ is averaged over all the grid positions such that final score of $D$ is obtained with respect to the input. For testing purposes, we directly use the averaged scores of $\hat{D}$ as a "detector" which is run over the grid to detect the abnormality from the input frame (see Sec. 2.2.2).

It is important to highlight that both $\{F_t\}$ and $\{O_t\}$ are here collected by using only the frames from*normal* scenarios in the identified zones provided by GP. The absence of abnormal events at the training phase makes it possible to train the discriminators corresponding to our two tasks without the need of supervised training data: $G$ acts as an implicit supervision for $D$. We hypothesize that the latter lies outside the discriminator's decision boundaries because they represent situations never observed during training and hence treated by $D$ as outliers. We use a *Bank of Discriminators* based on the identified zones provided by GP, which is grouped into two sets: $Set1$, which is trained on a straight path, and $Set2$ that is trained over the curves. The discriminator's learned decision boundaries can be used to detect unseen events as explained in the next section

### 2.2.2. *Anomaly detection*

Discriminators are used at the testing phase. More specifically, let $\hat{D}^{F\to O}$ and $\hat{D}^{O\to F}$ be the patch-based discriminators trained using the two channel-transformation tasks (see Sec. 2.2.1). Given a test frame $F$ and its corresponding optical-flow image $O$, we first produce the reconstructed $p_O$ and $p_F$ using $G^{F\to O}$ and $G^{O\to F}$, respectively. Then, the pairs of patch-based discriminators $\hat{D}^{F\to O}$ and $\hat{D}^{O\to F}$ are applied respectively to the first and second tasks. Such operation results in two scores for the ground truth observation: $S^O$ and $S^F$, and two scores for the prediction (reconstructed data): $S^{p_O}$ and $S^{p_F}$. The two scores are summed: $S_{observation} = S^O + S^F$, $S_{prediction} = S^{p_O} + S^{p_F}$, and the values in $S_{observation}$ and $S_{prediction}$ are normalized into the range $[0, 1]$. Note that, a possible abnormality in the observation (e.g., an unusual object/movement) corresponds to an outlier with respect to the data distribution learned by $\hat{D}^{F\to O}$ and $\hat{D}^{O\to F}$ during training. The presence of the anomaly results in a low value of $\hat{D}^{F\to O}(p_F, p_O)$ and $\hat{D}^{O\to F}(p_O, p_F)$ (prediction), but a high value of $\hat{D}^{F\to O}(F, O)$ and $\hat{D}^{O\to F}(O, F)$ (observation).

Hence, in order to decide whether an observation is abnormal with respect to the scores from the current bank of Discriminators, we simply measure the distance between predicted scores and observation scores such as shown in equation (4).

$$\tilde{Y} = S_{observation} - S_{prediction} \tag{4}$$

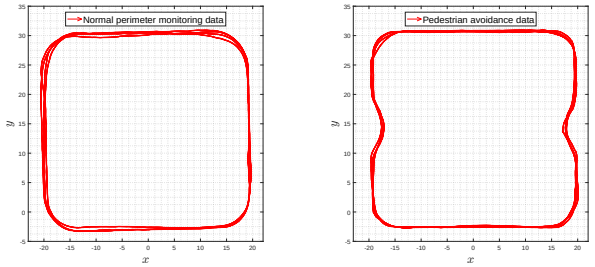
An error threshold $\tilde{Y}_{thres}$ is defined to detect the abnormal events: when $\tilde{Y}$ is higher than such threshold, the current agent's measurement is considered as an abnormal situation.

## 3. DATASET

Proposed approach is validated with data acquired from a real vehicle during a perimeter monitoring task. The 'iCab' vehicle is equipped with several heterogeneous sensors [16]. In this work, we consider data related to vehicles position and images grabbed from a frontal on-board camera Two different scenarios are considered, consisting of standard perimeter surveillance task and anomalies while executing it.

- Scenario 1: vehicle performs a rectangular path around the environment (perimeter monitoring), see Fig.3a.
- Scenario 2: vehicle performs an avoidance maneuver to avoid a static pedestrian and then continues the standard patrolling (Fig.4), see Fig.3b.

In both scenarios, the vehicle executes the correspondent task several times per each experiment.




(a) Perimeter monitoring (b) Pedestrian avoidance

**Fig. 3**: Observed tasks in the scene

**Fig. 4**: Pedestrian avoidance from 'iCab' on board camera

## 4. RESULTS

This section presents the results of proposed methods applied on the dataset presented in section 3. The normal behavior corresponds to perimeter monitoring defined in the Scenario 1.

### 4.1. Shared Level Self Awareness abnormality detection

As discussed in subsection 2.1, Fig. 7 shows the segmentation of GP into zones. In each zone, quasi-constant velocity models are valid. Large and small zones represent the action patterns for going straight and curving respectively. Additionally, each zone is used to create a KF valid in that specific area. As explained in subsection 2.1.1, by considering innovations generated by the bank of KFs based on the perimeter control task, it is possible to identify abnormalities simply by observing new trajectory data that does not correspond with the already characterized models. The value of $\xi_k$ measures the abnormality level at the time instant $k$. High innovation values indicate the presence possible anomalies in the scene. By processing position measurements from Scenario 2 (section 3) and analyzing innovations with respect to the normality model, it is possible to detect anomalies. The abnormality threshold $|\xi_{tresh}|$ is set at $0.4$ and

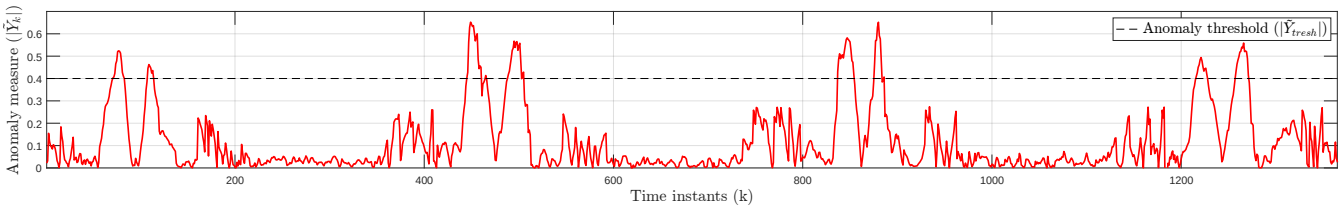



**Fig. 5**: SL anomaly measurements: perimeter control activity by GP through time with avoidance of static pedestrians.

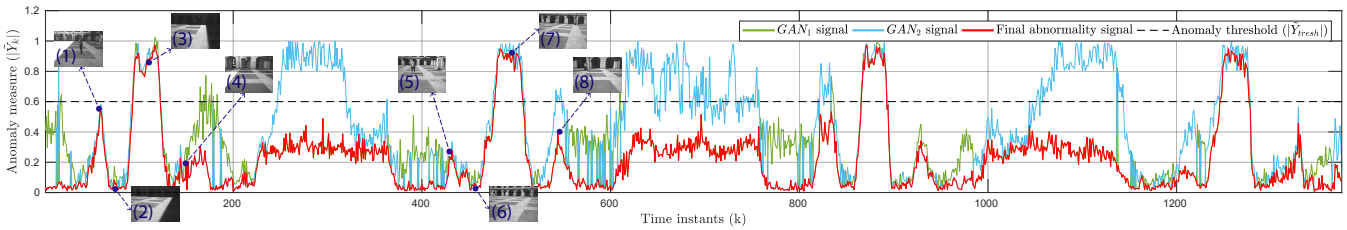



**Fig. 6**: PL anomaly measurements: the distances between the observations and predictions by GANs during the time.

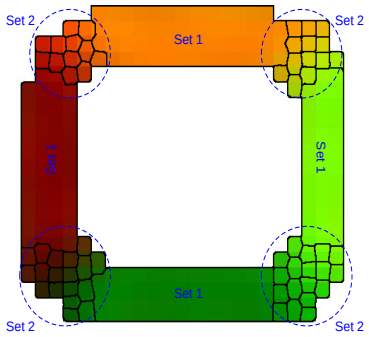



**Fig. 7**: Spatial information in terms of zones

produced anomaly detection results are shown in the Fig. 7. It is possible to find a pattern composed by two abnormal peaks that are associated to the avoidance of the standing pedestrian. An uniform anomaly pattern is not formed due to the straight component of the avoidance maneuver that follows the regular perimeter control behavior. In addition, under the threshold, other two behaviors can be recognized, i.e., straight and curve tracks performed by the vehicle. The lowest abnormality levels correspond to the straight parts of the track. Some abnormality peaks under the threshold value are created when the vehicle curves due to slightly different turning angles when performing the experiment laps.

### 4.2. Private Level Self Awareness abnormality detection

The bank of GANs are trained on the subsets of data based on GP zones. In our experiments, the bank of GANs is composed of two major subsets: $Set1$ and $Set2$, see Fig. 7. Each GAN detects the abnormality in the corresponding set on which is trained. The self-awareness model is tested on the second scenario discussed in subsection 3. Anomaly detection results associated to the PL, using the proposed bank of GANs are shown in Fig. 6. Three signals are shown in Fig. 6: The green and blue signals respectively show the computed signals by our $GAN_1$ (trained on $Set1$) and $GAN_2$ (trained on $Set2$). The red signal indicates the final abnormality measurement which is defined as the minimum value of $GAN_1$ and $GAN_2$. As it was expected, the obtained abnormality measurement in PL is aligned with SL results shown in Fig. 5.

Different parts of the curve can be associated and explained by considering the correspondent images acquired from the on-board sensor. Specifically, the small peak identified with number 1 can be justified by the presence of the pedestrian in the field of view of the camera: the vehicle do not start the avoidance maneuver yet, thus, it can be seen as a pre-alarm. The small peak in 1 corresponds to peak in 5, the latter is smaller due to the posture of the pedestrian, see correspondent images 1 and 7.The areas of the curve identified with numbers 2 and 3 or 6 and 7 correspond to the starting point of the abnormal maneuver and the avoiding behavior itself: it can be seen that peaks 3 and 7 are higher than the selected threshold and then correspond to an anomaly. After the small peak 4, that corresponds to the closing part of the avoidance turn, the vehicle goes back to the standard behavior. In particular, at this point of the curve, the vehicle is actually turning. In the wider area (from 220 to 380 secs.), the 'iCab' is moving straight. The slightly higher level of the abnormality curve in straight areas can be explained by a noise related to the vibration of the on-board camera due to the fast movement of the vehicle when increasing its speed.

It is notable that, the signal generated by $GAN_1$ becomes higher in the curving areas since it is only trained on $Set1$ for detecting straight paths. Similarly, the $GAN_2$ which is trained on $Set2$, generates higher scores on the straight path. However, both $GAN_1$, and $GAN_2$ can detect the abnormality area (pedestrian avoidance) where both generate a high abnormality score.

## 5. CONCLUSIONS

We presented a multi-perspective approach to detect anomalies for moving agents. Obtained results demonstrate the capability of our methodology to recognize anomalies using multiple viewpoints, namely PL and SL. A future research path could consist in combining information from different sources for decision making and robust the proposed self-awareness model. In particular, situational awareness and self-reactions could be increased with respect to the existing literature.

## 6. REFERENCES


[1] D. M. Ramík, C. Sabourin, R. Moreno, and K. Madani, "A machine learning based intelligent vision system for autonomous object detection and recognition," *Applied Intelligence*, vol. 40, no. 2, pp. 358–375, Mar 2014.

[2] D. Campo, A. Betancourt, L. Marcenaro, and C. Regazzoni, "Static force field representation of environments based on agents nonlinear motions," *Eurasip Journal on Advances in Signal Processing*, vol. 2017, no. 1, 2017.

[3] V. Bastani, L. Marcenaro, and C. S. Regazzoni, "Online nonparametric bayesian activity mining and analysis from surveillance video," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2089–2102, May 2016.

[4] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, Aug 2008.

[5] G. Pons-Moll, A. Baak, T. Helten, M. Mller, H. P. Seidel, and B. Rosenhahn, "Multisensor-fusion for 3d full-body human motion capture," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 663–670.

[6] Y. Charlon, W. Bourennane, F. Bettahar, and E. Campo, "Activity monitoring system for elderly in a context of smart home," *IRBM*, vol. 34, no. 1, pp. 60 – 63, 2013, Digital Technologies for Healthcare.

[7] E.B. Ermis, V. Saligrama, P.-M. Jodoin, and J. Konrad, "Abnormal behavior detection and behavior matching for networked cameras," 2008, cited By 15.

[8] R. Emonet, J. Varadarajan, and J.-M. Odobez, "Multi-camera open space human activity discovery for anomaly detection," 2011, pp. 218–223, cited By 13.

[9] K. Kim, D. Lee, and I. Essa, "Gaussian process regression flow for analysis of motion trajectories," 2011, pp. 1164–1171.

[10] D. Campo, M. Baydoun, and Cavallaro A. Regazzoni C. Marcenaro, L., "Modeling and classification of trajectories based on a gaussian process decomposition into discrete components," in *14th IEEE International conference on Advance Video and signal based surveillance*. 2017, AVSS 2017, IEEE Computer Society.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., pp. 2672–2680. Curran Associates, Inc., 2014.

[12] Z. Li and J. Chen, "Superpixel segmentation using linear spectral clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1356–1363.

[13] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV 2004, 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings*, 2004.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. S. Regazzoni, and N. Sebe, "Abnormal event detection in videos using generative adversarial nets," in *2016 IEEE International Conference on Image Processing, ICIP 2017 Beijing, China, September 17-20, 2017*, 2017.

[16] P. Marın-Plaza, J. Beltrán, A. Hussein, B. Musleh, D. Martın, A. de la Escalera, and J. M. Armingol, "Stereo vision-based local occupancy grid map for autonomous navigation in ros," in *11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications VISIGRAPP 2016*, 2016.