FORT-RAJ: A Hybrid Fisheye Model for Real-Time Pedestrian Trajectory Prediction

Yacine Amrouche, Sarra Bouzayane, Baptiste Magnier

FORT-RAJ: A Hybrid Fisheye Model for Real-Time Pedestrian Trajectory Prediction

Yacine Amrouche[†], Sarra Bouzayane[†] and Baptiste Magnier[‡,⋆]
[†] *Caplogy Innovation*, Vélizy, France
[‡] *EuroMov Digital Health in Motion, Univ Montpellier, IMT Mines Ales,* Alès, France
[⋆] *Service de Médecine Nucléaire, Centre Hospitalier Universitaire de Nîmes, Université de Montpellier*, Nîmes, France
y.amrouche@caplogy.com, s.bouzayane@caplogy.com, baptiste.magnier@mines-ales.fr

***Abstract*—This paper introduces FORT-RAJ, a hybrid model designed for pedestrian trajectory prediction in the context of top-view fisheye images. To achieve this, FORT-RAJ merges the FORT (Fisheye Online Realtime Tracking) algorithm, which tracks people using fisheye cameras without prediction capabilities, with the GATraj model, known for trajectory prediction but not yet adapted for fisheye images. The proposed method, FORT-RAJ, is designed to detect pedestrians, track their trajectories, and predict their future positions. It leverages the wide field of view of fisheye cameras while addressing the distortions inherent in such images. The experiments demonstrated that the FORT-RAJ model performs satisfactorily on fisheye images, achieving an Average Displacement Error (ADE) of 0.38 meters and an Final Displacement Error (FDE) of 0.42 meters.**

***Index Terms*—Trajectory prediction, Tracking models, Fisheye images, Distortions, Non-conventional sensors.**

I. Introduction

Reliable prediction of pedestrian trajectories is crucial to advances in emerging fields such as urban surveillance and intelligent transport systems. With the rise of autonomous vehicles and smart cities, the need to understand and anticipate human movement has never been more critical. Fisheye cameras, with their ability to capture wide panoramic views, have become invaluable tools in these applications. However, adapting existing trajectory prediction algorithms to the particularities of fisheye cameras remains a significant challenge.

The radial distortion in fisheye images warps not only objects but also the apparent trajectories of pedestrians, presenting significant challenges for models designed to operate in rectilinear perspectives [1]. Despite the proliferation of prediction techniques, much of the current research does not directly address the complications introduced by these types of cameras. Furthermore, converting fisheye data into a format that can be used by standard algorithms without losing relevant information about the actual trajectory is not straightforward and requires innovative approaches.

Nowadays, several algorithms are still attempting to predict object trajectories [2]–[4]. However, none exploit images from fisheye cameras, making them unsuitable for this task, and algorithms that do use fisheye images are limited to trajectory tracking without making predictions. [5] [6]. The aim of this paper is to combine these two advances to propose a model for predicting pedestrian trajectories adapted to these sensors.

The paper is organised as follows: the next Section II presents recent work on trajectory prediction models. Section III is devoted to the trajectory prediction models chosen to combine with the FORT tracking algorithm. Section IV presents the experiments and analyses the results. Section V concludes this work and outlines future prospects.

II. State of the art

Early attempts to model pedestrian movement approached crowd dynamics through analogies with physical systems. A notable example is the "Social Force Model" [7], which conceptualizes pedestrian movement using social forces analogous to physical forces. This model posits that pedestrians are influenced by attractive forces towards their destinations and repulsive forces to avoid collisions, thereby enabling the simulation of complex crowd movement dynamics with remarkable accuracy. With the advent of deep neural networks, trajectory prediction has significantly evolved, substantially influencing how visual and temporal data are processed to predict the movements of pedestrians and other moving objects. In this section, trajectory prediction models proposed for non-fisheye images are presented.

One of the early notable uses of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, in trajectory prediction was introduced by the "Social LSTM" model [8]. This model captures interactions between different agents in a shared space, emphasizing the importance of contextual data in trajectory prediction. Each pedestrian is represented by an LSTM cell that captures their movement state. These LSTM cells are connected to model the social interactions between pedestrians. The model aggregates the hidden states of each pedestrian's neighbors to form an enriched representation of the interactions. This combination of hidden states improves the accuracy of trajectory predictions.

Adversarial generation techniques, as illustrated by the "Social GAN" model [9], have enabled the generation of socially acceptable trajectories using Generative Adversarial Networks (GANs). In this model, each pedestrian is represented by a point, and the relative positional differences between pedestrians are used to model social interactions. This information is then integrated into the GAN, where a social pooling mechanism aggregates relevant information from neighbors, leading to trajectory prediction.

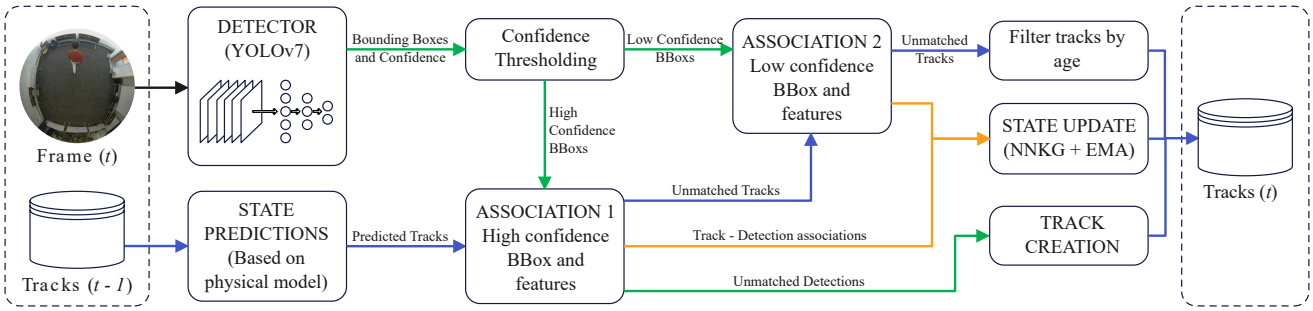

Fig. 1. Global architecture of the FORT algorithm for a frame at time t, see [6] for more details.

Model "SoPhie" [10] innovated by integrating attention mechanisms into GANs to predict paths that respect both social and physical constraints. This approach highlights the importance of considering both agent-agent and agent-environment interactions to improve trajectory prediction accuracy. SoPhie employs dual attention: social attention to capture interactions between pedestrians and physical attention to integrate environmental constraints. This dual attention allows the model to generate trajectories that are both socially acceptable and physically feasible.

The model NSP (Neural Social Physics) [11] combines the principles of social physics with neural networks to predict pedestrian trajectories. Each pedestrian in this model has a field of view in which the environment influences their movements. The field of view is determined by the current speed vector and is segmented into walkable and non-walkable areas. This alows so to present the areas of the environment that repel the pedestrian. This information is fed into the model to model social and physical interactions, providing more accurate predictions. A major difference between NSP and existing deep learning is the deterministic system built into NSP. Instead of learning a function linking input to output (as deep learning does as a black box), the deterministic system acts as a strong inductive bias and constrains the functional space in which the target match must lie. This is because a family of partial differential equations (PDEs) can be thought of as a flow connecting the input space and the output space, and learning is essentially a process of finding the PDE best suited to that flow. In addition to a better ability to fit the data, this strong inductive bias has two other advantages. Firstly, the learned model can help explain the movements, because the used PDE is a physical system where the learnable parameters have physical meanings. Secondly, after learning, the PDE can help to predict movements in very different scenes (e.g. with higher densities) and generate more plausible trajectories (e.g. with fewer collisions). This is difficult for existing deep learning because it requires significant extrapolation to unseen interactions between pedestrians.

The model Social-Transmotion (S-T) is based on a transforming approach to the prediction of human trajectories [12]. Using sophisticated attention mechanisms, this model can generate accurate and adaptive predictions based on the social and physical contexts of pedestrians. The model seamlessly integrates various types and quantities of visual cues, enhancing its adaptability to diverse data modalities and exploiting rich information for improved prediction performance. Its dual-transform architecture dynamically evaluates the importance of different visual cues from both primary and neighbouring pedestrians, effectively capturing relevant social interactions and body language cues. Indeed, this model explores the cues that humans consciously or unconsciously use to convey their mobility patterns. For example, individuals may turn their head and shoulders before changing direction while walking, a visual cue that cannot be captured using only a sequence of spatial locations in time. Similarly, social interactions can be anticipated through gestures such as hand signals or changes in head direction. To do this, the sequence of cues observed in the input is incorporated, along with the observed trajectories, to predict future trajectories.

By integrating social, physical, and contextual aspects, current methods not only improve prediction accuracy but also offer better adaptability to real-world scenarios. However, these models remain limited to conventional cameras and have yet to take advantage of the benefits offered by fisheye images.

III. A Hybrid Fisheye Model for Real-Time Pedestrian Trajectory Prediction

In this section, the contribution is presented in relation to existing work. The FORT model [6], designed for tracking people using a fisheye camera, has been integrated with the GATRAJ model [13], proposed for predicting trajectories from conventional images. The section begins with a detailed explanation of the two models, FORT and GATRAJ, before presenting the hybrid model, FORT-RAJ.

A. FORT: Fisheye Online Realtime Tracking

The FORT model was proposed in [6] with the aim of tracking people using Fisheye cameras. The Fig. 1 shows the architecture of the FORT model for an image at time t of the video. Detection Bounding Boxes (BBoxes) are created from the image and are divided into two groups according to the associated confidence. Tracks from previous images are updated with predicted new states and associated with high-confidence BBoxes. Remaining tracks are matched with low-confidence BBoxes. Unmatched tracks are filtered based on their age, and unmatched high-confidence BBoxes are initialized as new tracks. Finally, the Track-BBox associations are updated using a matched Kalman filter, and each track's characteristics are

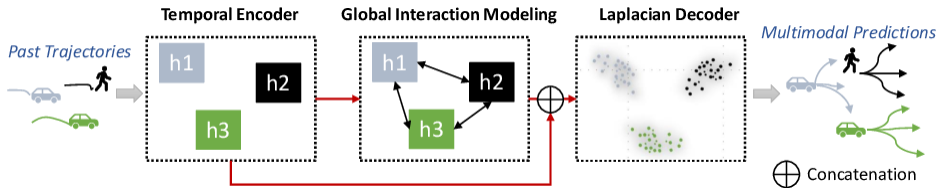

Fig. 2. GATraj: This model takes as input the observed trajectory of each agent and outputs multimodal predictions of their potential future trajectories [13].

refined using the Exponential Moving Average (EMA). The algorithm outputs objects called tracks, characterized by their state, attributes, and a unique identifier. Position and velocity information is stored in the state variable.

B. GATraj Model

The "GATraj" model [13] relies on a graph-based neural network and attention mechanisms to predict the trajectories of multiple agents. By integrating both spatial and temporal information, GATraj significantly improves the accuracy of predictions by taking into account the complex interactions between agents in dense environments. As shown in Fig. refgatraj, the GATraj framework consists of three parts:

- Temporal encoder: the temporal encoder takes as input the observed trajectory of each agent h_i. In Fig. 2, three agents' trajectories h_1, h_2, h_3 are respectively associated with vehicle (gray), human (black), and vehicle (green). This encoder learns temporal dynamic information from the observed trajectory to extract temporal dependencies over time. It should be noted that each agent's observed trajectory is processed independently, allowing their temporal dynamics to be handled in parallel.
- Global interaction : this module is used for modeling global interactions among all the concurrent agents in a given scene
- Laplace decoder: it is introduced to generate future multimodal trajectories, accounting for each agent's stochastic behavior. It takes as input the spatial–temporal dynamic context from the temporal encoder and the GCN module of the global interaction modeling. Its outputs are a set of different modes of the predicted trajectory distribution.

To summarize, the GATraj model uses attention layers to capture the relationships between different agents, assigning weights to connections based on their relevance. This architecture enables interactions between pedestrians to be modelled more effectively, taking into account distance and temporal relationships. The information from each agent is passed through several layers of graphs, providing an in-depth understanding of social dynamics and movements in space as shown in the diagram of Fig 2. The output of this model are the multimodal predictions of their possible future trajectories.

C. FORT-RAJ:

The proposed hybrid model, named FORT-RAJ, combines the FORT and GATraj algorithms to detect, track, and predict pedestrian trajectories using fisheye images 3.

The FORT component begins by capturing a fisheye image, which is then processed by the YOLOv7 [14] algorithm to detect the pedestrians present in the scene. The detected pedestrians' features are extracted using ResNeXt-50, enabling their re-identification across multiple frames. They are tracked using an enhanced Kalman filter and a neural network (NN) for Gain calculation. Then, pedestrian trajectories are updated with each new detection using a combination of Kalman filter and exponential moving average (EMA). Finally, new detections are associated with existing tracks using Intersection over Union (IoU) and re-identification (ReID) techniques, ensuring that each pedestrian maintains a unique identity across frames. The updated pedestrian trajectories provided by FORT are subsequently fed into the GATraj model.

The GATraj component begins by extracting temporal information from the tracked trajectories using a temporal encoder that combines Conv1D (ie. layer used in neural networks to apply a one-dimensional convolution operation), MLP (ie. Multilayer Perceptron), Self-Attention (ie. a mechanism used in neural networks to dynamically weigh the importance of different parts of the input data), and LSTM (Long Short-Term Memory). This captures the temporal and contextual dynamics of pedestrian movements. Next, the Global Interaction module, based on graph convolutional networks (GCN), models the global interactions between pedestrians, considering the collective dynamics and mutual influences of agents in the scene. The decoder then predicts the multi-modal future trajectories of pedestrians, accounting for multiple movement possibilities based on observed dynamics and predicted interactions. The predicted multi-modal trajectories of pedestrians are output, offering a prospective view of likely pedestrian movements in the scene. The selection of the GATraj model, along with the performance of the new FORT-RAJ model, will be discussed in the following section.

IV. Experiments and Results

In this section, the datasets used for the experiments are presented, and the evaluation metrics commonly used in the literature are discussed, which will be employed to assess the models. Next, the choice of the GATraj model for trajectory prediction is justified. Finally, the FORT-RAJ model is evaluated, and the results are discussed.

A. Datasets

Experiments are conducted on two well-established non-fisheye datasets: ETH [15] and UCY [16]. These datasets contain annotated trajectories of multiple pedestrians in real scenes, including rich social interactions and are the main benchmarks for pedestrian trajectory prediction. The third dataset is the Fisheye HABBOF[1]. It includes annotated bound-

[1]https://vip.bu.edu/projects/vsns/cossy/datasets/habbof/

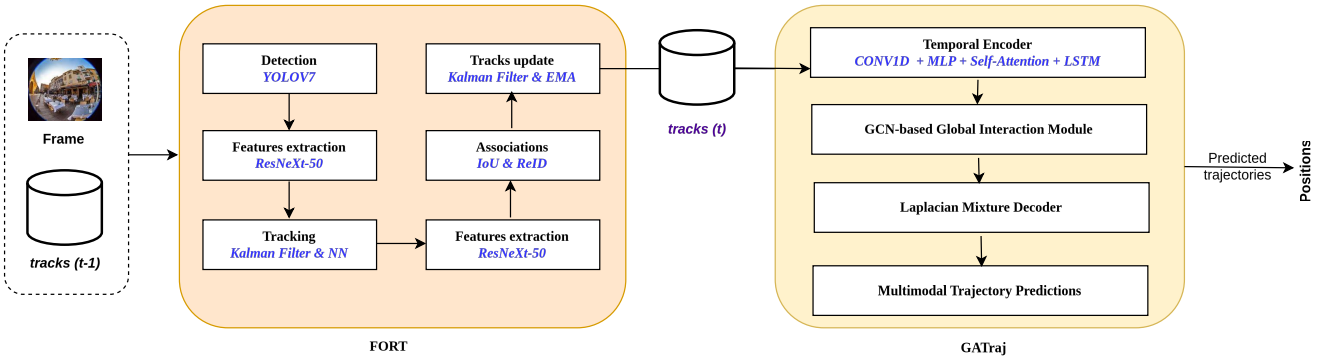

Fig. 3. FORT-RAJ : this framework takes as input tracks from the previous frames and output the potential future positions.

ing boxes, but assigns the annotation "person" to all detected individuals without providing unique identifiers.

- ETH dataset is used for pedestrian detection and consists of a test set with 1,804 images across three video clips. The data is captured from a car-mounted stereo system with a resolution of 640×480 (bayered) and a frame rate of 13 to 14 FPS (frames per second). The ETH dataset includes two scenes: ETH and HOTEL.
- UCY dataset features real pedestrian trajectories in scenarios rich with multiple human interactions, captured at 2.5 Hz. It comprises three sequences (Zara01, Zara02, and UCY), recorded from a top-down view in public spaces.
- HABBOF (Human-Aligned Bounding Boxes from Overhead Fisheye Cameras) is a fisheye dataset including four scenarios: 2 capture meetings in a conference room with up to three persons, containing 1,119 and 1,121 frames respectively; and 2 depict computer laboratory scenarios with up to four persons, containing 1,792 and 1,805 frames.

The ETH and UCY datasets contain real pedestrian trajectories with scenarios rich in multi-human interactions.

B. Evaluation metrics

Evaluation metrics and data sets play a crucial role in the development and validation of these algorithms. In standard trajectory prediction, metrics such as mean square error (MSE) and accuracy can be computed. However, these metrics may not fully capture the complexity of trajectories in fisheye images. For this reason, the two metrics ADE and FDE will be used to evaluate our models.

- Average Displacement Error (ADE) represents the average L_2 distance between the actual and predicted trajectories at each instant in the prediction window. For M trajectories, the ADE is calculated as follows:

$$\text{ADE} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}\sqrt{(x_{ij}-\hat{x}_{ij})^2+(y_{ij}-\hat{y}_{ij})^2},$$

  where:
  - M denotes the total number of trajectories,
  - N_j is the number of points in the jth trajectory,
  - (x_{ij}, y_{ij}) represent the actual coordinates at point i of the jth trajectory,
  - $(\hat{x}_{ij}, \hat{y}_{ij})$ are the predicted coordinates at point i of the jth trajectory.

- Final Displacement Error (FDE) is a measure of the error between the predicted final position and the actual final position. For M trajectories, the FDE is calculated as:

$$\text{FDE} = \frac{1}{M}\sum_{j=1}^{M}\sqrt{(x_{N_jj}-\hat{x}_{N_jj})^2+(y_{N_jj}-\hat{y}_{N_jj})^2},$$

  where (x_{N_jj}, y_{N_jj}) represent the actual coordinates at the end of the j-th trajectory and $(\hat{x}_{N_jj}, \hat{y}_{N_jj})$ the predicted coordinates at the end of the j-th trajectory respectively.

These metrics are used to evaluate the models presented above on the selected ETH [15] and UCY [16] databases.

C. Experiments and analysis

Pedestrian trajectory prediction models have mainly been developed for perspectives images, as previously mentioned. The ETH and UCY datasets are also acquired from these sensor types. Habbof, however, is a fisheye Dataset. Our initial aim is therefore to test the models NSP, ST and GATraj on these NON-fisheye datasets in order to determine which performs best and to gain a better understanding of their behaviour. Then, the proposed model FORT-RAJ is assessed on a fisheye dataset to select and adapt the model that best complements FORT for fisheye images.

1) Model Selection: The results showed that the GATRAJ model performed best in predicting trajectories, slightly outperforming the NSP model and significantly outperforming the ST model (cf. Table I). These results are preliminary. In our future research, we plan to discuss and test several other models to strengthen our study, including the GAN, LSTM, GCN and CRF models algorithms which are designed for conventional cameras. To our knowledge, there is no existing model adapted to fisheye lenses for trajectory prediction, hence the innovation of our FORT-RAJ model.

2) FORT-RAJ evaluation: The figure 4 shows a concrete example of captures taken at seconds 32, 36, 45 and 56 using the FORT model on our own dataset. We observe that FORT successfully tracked the person moving in the room while maintaining a consistent identifier. This output is then used as input into GATraj for prediction.

To evaluate FORT-RAJ model, which is designed to predict a person's trajectory based on fisheye images, we first annotated the Lab1 scenario of the HABBOF dataset with unique

TABLE I
COMPARING THE PERFORMANCE OF TRAJECTORY PREDICTION MODELS (ERROR EXPRESSED IN METERS).

Model \ Database	**ETH** ADE/FDE	**HOTEL** ADE/FDE	**UNIV** ADE/FDE	**Zara01** ADE/FDE	**Zara02** ADE/FDE	**Average** ADE/FDE	**Mean** (ADE+FDE)/2
GATraj	0.26/0.42	0.10/0.15	0.21/0.38	0.16/0.28	0.12/0.21	0.17/0.29	0.23
NSP	0.25/0.24	0.09/0.13	0.21/0.38	0.16/0.27	0.12/0.20	0.41/0.81	0.61
S-T	0.93/1.81	0.32/0.60	0.54/1.16	0.42/0.90	0.32/0.7	0.51/1.03	0.77

identifiers and then created our own test dataset. Although HABBOF was initially annotated, it lacked identifiers. Our dataset consists of a video composed of 921 frames (25 fps), created using our own fisheye camera and manually annotated. This video features a single person moving in a room. These frames are input into FORT, which performed very well in providing the person's tracks as output. These tracks are then input into GATraj to predict the next positions. FORT-RAJ achieved satisfactory results, with ADE = **0.38**m (meters) and FDE = **0.42**m. The next steps involve adapting the GATraj model to fisheye images and testing on additional datasets to optimize the results. Other models will then be evaluated and compared with FORT-RAJ in real-time under varied and dynamic conditions.

V. CONCLUSION

This paper proposes a hybrid model, FORT-RAJ, for predicting pedestrian trajectories using fisheye cameras. FORT-RAJ combines the FORT (Fisheye Online Realtime Tracking) model, which tracks pedestrians based on fisheye images, with the GATraj model, which predicts trajectories but is not adapted to fisheye images. The GATraj model was selected following a comparative study with the S-T (Social Transmotion) and NSP (Neural Social Physics) models on non-fisheye datasets. FORT-RAJ was then tested on a fisheye dataset and yielded promising results, with an ADE of 0.38 meters and an FDE of 0.42 meters. Our future plans involve testing other prediction models, such as Social-LSTM, Social-GAN, and SoPhie, and integrating them with FORT. The resulting models will be evaluated on more complex fisheye datasets.

REFERENCES

[1] Sistu, G. & Yogamani, S. FisheyeDetNet: 360° Surround view Fisheye Camera based Object Detection System for Autonomous Driving. (2024)

[2] Chib, P. & Singh, P. LG-Traj: LLM Guided Pedestrian Trajectory Prediction. *ArXiv Preprint ArXiv:2403.08032*. (2024)

[3] Yang, J., Chen, Y., Du, S., Chen, B. & Principe, J. IA-LSTM: interaction-aware LSTM for pedestrian trajectory prediction. *IEEE Transactions On Cybernetics*. (2024)

[4] Dong, Y., Wang, L., Zhou, S., Hua, G. & Sun, C. Recurrent Aligned Network for Generalized Pedestrian Trajectory Prediction. *ArXiv Preprint ArXiv:2403.05810*. (2024)

[5] Haggui, O., Bayd, H. & Magnier, B. Centroid human tracking via oriented detection in overhead fisheye sequences. *The Visual Computer*. **40**, pp. 407–425 (2024)

[6] Odic, N., Faure, B. & Magnier, B. FORT: Fisheye Online Realtime Tracking with an Improved Kalman Filter. *IEEE MMSP*. pp. 1-6 (2023)

[7] Helbing, D. & Molnar, P. Social force model for pedestrian dynamics. *Physical Review E*. **51**, 4282 (1995)

[8] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L. & Savarese, S. Social lstm: Human trajectory prediction in crowded spaces. *IEEE CVPR*. pp. 961-971 (2016)

[9] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S. & Alahi, A. Social gan: Socially acceptable trajectories with generative adversarial networks. *IEEE CVPR*. pp. 2255-2264 (2018)

[10] Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H. & Savarese, S. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. *IEEE/CVF CVPR*. pp. 1349-1358 (2019)

[11] Yue, J., Manocha, D. & Wang, H. Human trajectory prediction via neural social physics. *ECCV*. pp. 376-394 (2022)

[12] Saadatnejad, S., Gao, Y., Messaoud, K. & Alahi, A. Social-Transmotion: Promptable Human Trajectory Prediction. *ArXiv Preprint ArXiv:2312.16168*. (2023)

[13] Cheng, H., Liu, M., Chen, L., Broszio, H., Sester, M. & Yang, M. GATraj: A graph- and attention-based multi-agent trajectory prediction model. *ISPRS Journal Of Photogrammetry And Remote Sensing*. **205** pp. 163-175 (2023)

[14] Wang, C., Bochkovskiy, A. & Liao, H. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *IEEE/CVF CVPR*. pp. 7464-7475 (2023)

[15] Pellegrini, S., Ess, A., Schindler, K. & Van Gool, L. You'll never walk alone: Modeling social behavior for multi-target tracking. *IEEE ICCV*. pp. 261-268 (2009)

[16] Lerner, A., Chrysanthou, Y. & Lischinski, D. Crowds by example. *Computer Graphics Forum*. **26**, 655-664 (2007)

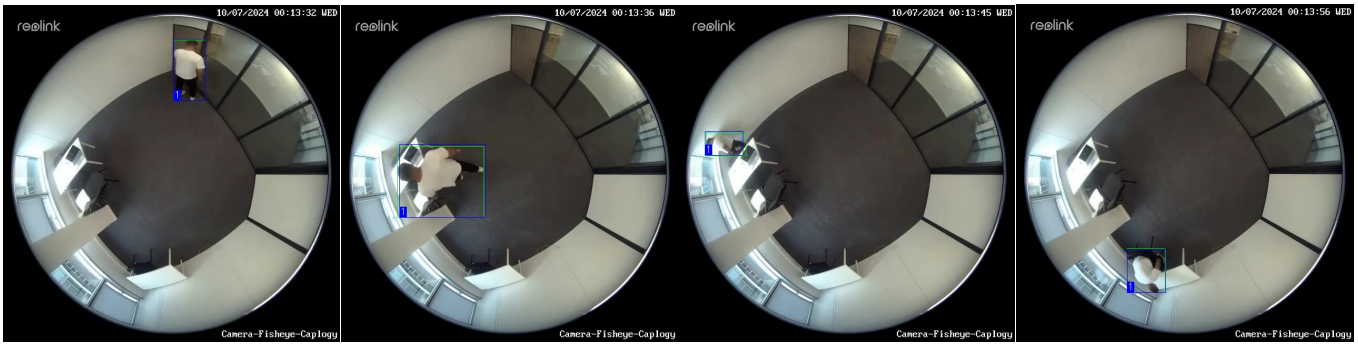

Fig. 4. Tracking Person Movement with FORT Model testesd on our dataset : Selected Frames in seconds (32s, 36s, 45s and 56s).