Trajectory Prediction in First-Person Video: Utilizing a Pre-Trained Bird's-Eye View Model

Masashi Hatano, Ryo Hachiuma and Hideo Saito
Graduate School of Science and Technology, Keio University, Yokohama, Japan

Keywords: Trajectory Prediction, Egocentric Video.

Abstract: In recent years, much attention has been paid to the prediction of pedestrian trajectories, as they are one of the key factors for a better society, such as automatic driving, a guide for blind people, and social robots interacting with humans. To tackle this task, many methods have been proposed but few are from the first-person perspective because of the lack of a publicly available dataset. Therefore, we propose a method that uses egocentric vision, which does not need to be trained with a first-person video dataset. We made it possible to utilize existing methods, which predict from a bird's-eye view. In addition, we propose a novel way to consider semantic information without changing the shape of the input to apply to all existing bird's-eye methods that use only past trajectories. Therefore, there is no need to create a new dataset from egocentric vision. The experimental results demonstrate that the proposed method makes it possible to predict from an egocentric view via existing methods of bird's-eye view. The proposed method qualitatively improves trajectory predictions without aggravating quantitative accracy, and the effectiveness of predicting the trajectories of multiple people simultaneously.

1 INTRODUCTION

Although future trajectory prediction is a challenging task, it is one of the key factors for a better society, such as automatic driving, a guide for blind people, and social robots interacting with humans. This task has been attracted many researchers (Bhattacharyya et al., 2019; Gupta et al., 2018; Sadeghian et al., 2019; Deo and Trivedi, 2020; Lee et al., 2017; Liang et al., 2020; Mangalam et al., 2020; Zhao et al., 2019). See (Rudenko et al., 2020) for an overview of future trajectory prediction. Various solutions from different perspectives have been proposed; however, few have tried to predict the future location of pedestrians from egocentric vision because the field of egocentric future prediction is very young and still developing, as mentioned in (Rodin et al., 2021).

The existing methods for predicting the trajectories of pedestrians can roughly be divided into two main categories: a third-person viewpoint or bird's-eye view (Alahi et al., 2016; Gupta et al., 2018; Zhang et al., 2019; Pang et al., 2021), and the other predicts trajectories from a first-person perspective (Yagi et al., 2018; Qiu et al., 2021; Huynh and Alaghband, 2020). While the former method can accurately use past trajectories in the world coordinate system, which are the most important information for future trajectories, as inputs, it cannot use detailed information such as a pedestrian's posture state. In addition, the former method is, in fact, impractical in terms of the ubiquitousness and availability of surveillance cameras, which provide images from the third-person perspective. The latter method can use detailed information such as human poses; however, models predicting trajectories from a first-person perspective must be trained with an egocentric vision dataset. To the best of our knowledge, there are few publicly available datasets of trajectory predictions from egocentric vision as first-person video contains private information such as pedestrians' faces. In our work, we made it possible to utilize existing methods of bird's-eye coordinates to apply to the egocentric view, which does not require the creation of a dataset or training model.

Moreover, none of the previous work (Liang et al., 2019; Kosaraju et al., 2019; Sadeghian et al., 2019; Salzmann et al., 2020) tried to change the shape of the input to adapt to the model structure. Researchers changed the structure of the model instead, as it is easy to extract information and integrate it with the prediction model. Nonetheless, this method of considering semantic information requires training with a dataset. There is a publicly available dataset for bird's-eye coordinate models; therefore, the segmen-

Hatano, M., Hachiuma, R. and Saito, H.
Trajectory Prediction in First-Person Video: Utilizing a Pre-Trained Bird's-Eye View Model.
DOI: 10.5220/0011683300003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 624-630
ISBN: 978-989-758-634-7; ISSN: 2184-4321

tation prediction model can be obtained without a problem. In contrast, the model from egocentric vision has a problem in terms of the dataset. We addressed this issue by proposing a novel method of considering semantic information that transforms the boundary information, obtained from the first-person viewpoint image, into the same shape as the inputs of the existing model, predicting from a bird's-eye view. This leads to unnecessary training with the dataset from a first-person perspective, but it enhances the prediction qualitatively while maintaining quantitative accuracy.

To demonstrate the effectiveness of the proposed method, three main comparisons were conducted. The first compares prediction accuracy with and without the homographic transformation, the second is a comparison with and without the terrain information, and the third comparison is for with and without using SP (Social Pooling). Regarding the socially acceptable predictor to make these comparisons, the pre-trained model of SocialGAN (Gupta et al., 2018) was used.

In this paper, there are mainly three contributions.

- We made it possible to apply existing methods that use bird's-eye coordinates as input to make forecast them from a first-person viewpoint. We can benefit from multi-person prediction from existing methods that considers human-interaction.
- We came up with a novel way of taking semantic information into account, which applies to all socially acceptable trajectory predictors, using only the past trajectory, resulting in qualitatively improving the trajectory prediction.
- We realized the proposed method without training and fine-tuning with an egocentric vision dataset.

2 RELATED WORK

Trajectory prediction can be used for a variety of purposes. One is Autonomous driving. Autonomous driving must predict the trajectories of pedestrians to avoid colliding with them. Recently researchers (Cai et al., 2022; Rasouli et al., 2019; Marchetti et al., 2020; Poibrenski et al., 2020; Makansi et al., 2020) have been tackling this issue, and most methods predict from a driving car's camera as they would like to apply the methods to intelligent driving. In this domain, a significant number of approaches have been introduced, as many publicly available datasets have been provided. In contrast, in this paper, we focus on predicting from a head-mounted camera of a pedestrian because the objective is to apply to assistive technologies and social robots that interact with human. This task, which we would like to address, is much more challenging than the one in autonomous driving because few public datasets are provided, and head-mounted cameras are unsteady compared to driving cameras.

Turning to human trajectory prediction from the first-person viewpoint, few researchers (Yagi et al., 2018; Qiu et al., 2021; Huynh and Alaghband, 2020) have addressed this problem, which aims at socially acceptable and efficient trajectory navigation. future person localization (FPL) (Yagi et al., 2018) is the first work to predict the future locations of people (not the camera-wearer) in egocentric videos from wearable cameras. Then, indoor future person localization (IFPL) (Qiu et al., 2021), which is the latest work in this task, was introduced to adapt for indoor scenes. Both methods take detailed information such as human body pose into account and predict points of future locations or the entire bounding box for pedestrians. Training an end-to-end model for trajectory prediction from an egocentric view is difficult due to the lack of a dataset. Both sets of authors created datasets and provide them on request, but these datasets are less diverse in terms of scene diversity and scale than existing datasets in areas such as automated driving. However, the proposed method solves these issues by performing coordinate transformations using homography. This transformation enables us to use existing methods, with only the past trajectory from a bird's-eye viewpoint; therefore, there is no need to train a model with first-person videos, which creates privacy issues.

Many more researchers have been attracted by predicting from a third-person viewpoint or bird's-eye viewpoint. The approach can be divided into two types: using rich information such as a semantic map of the image (Liang et al., 2019; Kosaraju et al., 2019; Sadeghian et al., 2019; Salzmann et al., 2020) and using only the past trajectory (Gupta et al., 2018; Alahi et al., 2016; Pang et al., 2021; Amirian et al., 2019; Choi and Dariush, 2019; Zhang et al., 2019; Katyal et al., 2020). Most of the previous work uses LSTM-based (Hochreiter and Schmidhuber, 1997) or Generative Adversarial Network-based (Goodfellow et al., 2014) models. Using a third-person or bird's-eye viewpoint is helpful as they provide accurate past trajectories in world coordinates without missing information caused by the occlusion. Nevertheless, they are impractical due to the availability and ubiquitousness of bird's-eye viewpoint videos. In contrast, although the proposed method is based on these third-person viewpoint predictors, it predicts from an egocentric view; therefore it is, in fact, practical. As for

the consideration of semantic information, we developed the novel idea of considering boundary information.

3 METHOD

3.1 Problem Formulation

Let $p_i^t \in \mathbb{R}^2$ denote the position of pedestrian i at time t in a single image frame where there are n pedestrians in total. The past trajectory of a pedestrian i is $p_i = \{p_i^t, t = 1,2,...,t_{past}\}$, and $P = \{p_i, i = 1,2,...,n\}$ represents the past trajectories of all pedestrians in a scene. Similar to this representation, the future trajectory of pedestrian i at time t is denoted as q_i^t. $q_i = \{q_i^t, t = t_{past}+1,...,t_{pred}\}$ and $Q = \{q_i, i = 1,2,...,n\}$ represent the future trajectory of pedestrian i and all future trajectories, respectively. In addition, $\hat{q}_i = \left\{\hat{q}_i^t, t = t_{past}+1,...,t_{pred}\right\}$ and $\hat{Q} = \{\hat{q}_i, i = 1,2,...,n\}$ denote the estimated future trajectory of pedestrian i and all estimated future trajectories in a scene with the existing socially acceptable trajectory prediction model f_θ, using only the past trajectories of all pedestrians:

$$\hat{Q} = f_\theta(P), \tag{1}$$

where θ is the parameters of a pre-trained trajectory prediction model. In this work, θ remains unchanged.

In addition, $e_i^t \in \mathbb{R}^2$ denotes the position of a fictitious pedestrian (defined in Section 3.3) i at time t, which is the position of an extracted point depicted in a blue rectangle in Figure 1. The past trajectory of a fictitious pedestrian i is $e_i = \{e_i^t, t = 1,2,...,t_{past}\}$ and $E = \{e_i, i = 1,2,...,m\}$ represents the past trajectories of all fictitious pedestrians in a scene where there are m extracted points in total. As the fictitious pedestrians are assumed to be at a standstill, the well-formed formula

$$\forall i, \exists c \in \mathbb{R}^2, \forall t, e_i^t = c, \tag{2}$$

is established, where c is a two-dimensional bird's-eye coordinate. Then, the estimated future trajectories that consider semantic segmentation information are

$$\hat{Q} = f_\theta(P,E). \tag{3}$$

In the following section, we detail the part of coordinate transformation and boundary extraction via a semantic segmentation map. As our objective is to utilize an existing pre-trained model without additional training because of the lack of first-person videos for trajectory prediction, the explanation of the trajectory prediction model is not given.

3.2 Coordinate Transformation

Transforming the coordinates is important in the proposed method because existing methods, which use bird's-eye coordinates, cannot be applied to infer from the first-person perspective without this technology, as proved in Section 4.2. One of the advantages of using methods that predict from the world coordinate system is that past trajectories are well considered.

Coordinate transformation is applied to pedestrians' foot coordinates and the extracted points using semantic segmentation, as shown in Figure 1. As can be seen in Figure 1, for each image frame, object detection (Carion et al., 2020; Redmon et al., 2016) is applied to detect the pedestrians in the image, and then the center of the two bottom corners of the bounding box is regarded as each detected person's foot coordinate in the screen coordinate system. After the screen coordinates are found, Equation 4 is applied to transform the coordinates from an image screen system to a world system:

$$[\mathbf{R}|\mathbf{t}]^{-1}\mathbf{K}^{-1}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{4}$$

where $\mathbf{K}$ is the intrinsic matrix of the camera, $\mathbf{R}$ is the rotation matrix, $\mathbf{t}$ is the translation vector, u and v are two-dimensional coordinates in the image, and X, Y, and Z are three-dimensional coordinates in the world. The height, Y, in the world coordinate system is unnecessary for the input data as we use the trajectory predictor, forecasting from a bird's-eye view.

By performing this transformation for each image frame with the corresponding camera parameters, the input data for trajectory prediction is prepared. In the same way, the extracted points, which are detailed in Section 3.3, are also transformed to bird's-eye coordinates. This is done every d frame, unlike the detected pedestrians, as a single image provides boundary information within 15 m from the camera-wearer.

3.3 Boundary Extraction

Normally, semantic segmentation (Xie et al., 2021; Yuan et al., 2020) is integrated with the prediction module when the semantic information should be considered; however, in the proposed method, semantic information is obtained by converting the inference results of semantic segmentation to the same input as the trajectory predictor. This is shown in Figure 2. This aims to avoid pedestrians' trajectories from being in inappropriate areas such as outside the road or on the wall. To do that, we regarded extracted bound-

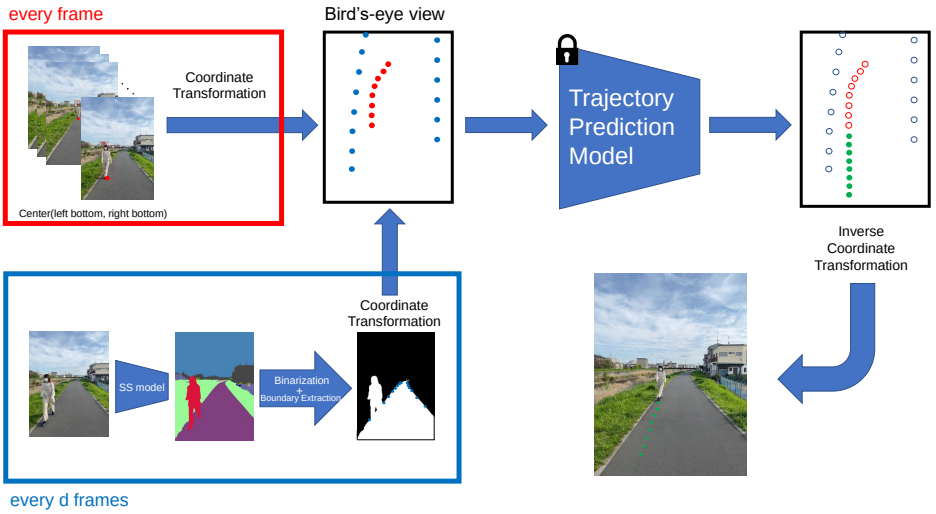

Figure 1: An overview of the proposed method for a single agent. The part surrounded by a red rectangle at the top left of the image shows how to transform the human past trajectories into bird's-eye coordinates. The part surrounded by a blue rectangle at the bottom left of the image shows how the border information is transformed and compressed to the bird's-eye coordinates in the proposed method. (SS stands for semantic segmentation).

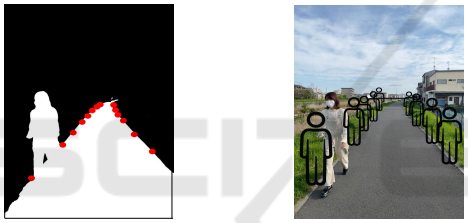

Figure 2: The concept of the proposed consideration of boundary information. The boundary information is obtained by regarding each extracted point as a standing fictitious pedestrian.

ary points, which are between inappropriate areas or not, as standing pedestrians. We call them fictitious pedestrians because there are no pedestrians on the boundary in the real world. As the existing trajectory predictor we used considers the avoidance of collision between pedestrians, the predicted trajectories will be in the appropriate area by using these fictitious standing pedestrians.

As can be seen in the blue rectangle in Figure 1, a first-person view image goes through semantic segmentation neural networks to be segmented into several classes, including roads and sidewalks. After obtaining the semantic segmented image, a mask image, with which it is determined whether the class is a walkable area or not, is generated. In our case, road and sidewalk classes are walkable areas, and the others are unwalkable areas. Then, many points of the boundary between prohibited and permitted areas are extracted from the masked image, and each point is transformed into the world coordinate system and given a specific identification number. In this way, boundary information is transformed into the same shape as the input of simple trajectory predictors such as SocialGAN.

As the trajectory prediction method, we used the same method as SocialGAN: 3.2 s for observation and the same amount of time for future prediction. A one-time step is defined as 0.4 s; therefore, there are eight time steps for observation and prediction, resulting in the need for 16 time steps to evaluate the predicted trajectory.

4 EXPERIMENTS

4.1 Overview

To prove the effectiveness of the proposed method, we compared the prediction accuracy with and without coordinate transformation, with and without consideration of boundary information, and with and without consideration of social interaction. The first and third comparisons were made with the metrics of two major trajectory prediction accuracy: average displacement error (ADE) and final displacement error (FDE), whereas the second comparison was made with the number of times that a pedestrian goes in a prohibited area, as well as the two distance metrics.

For the dataset to perform these experiments, we collected it ourselves, using an application that provides image sequence and camera parameters via

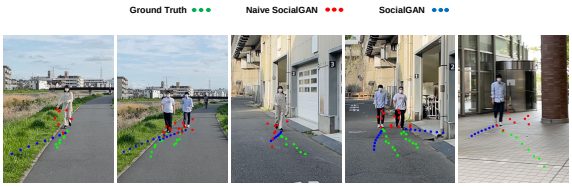

Figure 3: Qualitative results of trajectory predictions with and without coordinate transformation.

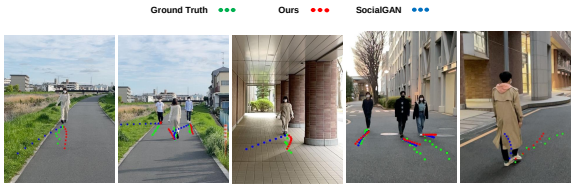

Figure 4: Qualitative results of trajectory predictions with and without the addition of fictitious pedestrians that are on the boundaries between appropriate and inappropriate areas.

ARKit[1]. The dataset has a variety of scenes in terms of the number of pedestrians and the type of prohibited area around pedestrians, such as bushes next to roads or buildings walls. This dataset contains 10 different scenes and around 100 input sequences in total.

4.2 Necessity of Coordinate Transformation

To accurately transform the coordinates from the screen system, to the world system, we utilized the intrinsic and external matrices provided by ARKit. As shown in (Seiskari et al., 2022), the performance of ARKit is comparable to the state of the art; thus, it enables us to get precise camera parameters. However, even if ARKit provides precise intrinsic, rotation, and translation matrices, it is difficult to get the exact world coordinates at points that are far from the camera-wearer. The transformation becomes worse and less accurate when the object points are more than 15 m from the camera-wearer. Although ARKit has this problem, which has a negative effect on the input data for trajectory prediction, the transformation is almost perfect if we use object points within 15 m of the camera.

To make this comparison, the difference is two-dimensional coordinates of input, and there are no differences in terms of the model or dataset. For the non-coordinate transformation, the pedestrian's foot coordinate in the image frame was used as the input for the trajectory predictor. The results are summarized in Table 1, and qualitative results are shown in Figure 3.

As Table 1 indicates, the average and final errors are huge if we do not apply the coordinate transformation. The trajectory prediction without coordinate transformation is inconsistent because the sequence of two-dimensional coordinates also lacks consistency, as can be seen in Figure 3. Each coordinate in the image depends on where the camera-wearer is and which direction the camera faces. In this sense, coordinate transformation, from the screen coordinate system to the world coordinate system, helps to maintain consistency between image frames, as the world coordinates are independent of the translation or rotation matrices of the camera.

[1]url:https://developer.apple.com/documentation/arkit/

4.3 Effectiveness of Distributing Fictitious Pedestrians

Similar to the previous experiment, the comparison between with and without using boundary information was performed in the same environment except for the input of the trajectory predictor. The input with and without the boundary information has the same input in terms of real pedestrians, but the former input contains the world coordinates of fictitious pedestrians as well as those of non-fictitious pedestrians. The results are summarized in Table 1, in which how many times pedestrians were predicted to penetrate prohibited areas (we call the percentage of this figure "area error") is shown in addition to the quantitative metrics, and qualitative results are shown in Figure 4.

As can be seen in Table 2, both quantitative metrics, ADE and FDE, are neither improved nor degraded; however, the area score is substantially enhanced with the addition of the boundary information to the input. It helps reduce the number of times pedestrians are forecast to be in the inappropriate area by half.

4.4 Effectiveness of Predicting Multiple Pedestrians Simultaneously

To show the effectiveness of considering multiple pedestrians at once, we compared between with and without social pooling in the prediction model. As can be seen in Table 1, ADE and FDE are improved if we consider the social interaction in a scene at once. In addition, as shown in Figure 5, the model that considers social interaction improves trajectory prediction qualitatively: A collision can be observed if we do not use social pooling. Social pooling was invented several years ago; however, most of the previous work of first-person perspective (Yagi et al., 2018;

Table 1: Results for the comparison in terms of three metrics: ADE, FDE, and area error.

	Naive Social-GAN	SocialGAN w/o SP	SocialGAN(Gupta et al., 2018)	Ours
ADE	27.22	10.65	9.07	8.93
FDE	50.08	13.12	9.34	9.33
area error	-	-	0.417	0.202

Table 2: Comparison of the techniques for each method.

	Naive SocialGAN	SocialGAN w/o SP	SocialGAN	Ours
Coordinate transformation		✓	✓	✓
Social pooling			✓	✓
Boundary information				✓

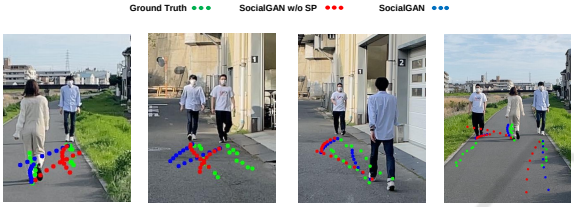

Figure 5: Qualitative results for trajectory predictions with and without social pooling.

Qiu et al., 2021) does not consider social interaction because they created the dataset on their own. However, the annotation was done by only one pedestrian in a scene, even if there are several pedestrians. In this sense, the proposed method, which utilizes existing methods, is much less expensive in terms of cost, as we realized multiple pedestrians prediction without high annotation costs.

5 CONCLUSIONS

In this work, we present a method that takes egocentric view video as input but applies the existing process, predicting from a bird's-eye coordinate to avoid being hindered by privacy issues and to utilize the current fairly good predictors. This approach improves the drawbacks of the first and third-person perspectives: the huge cost of creating an egocentric vision dataset, and the lack of ubiquitousness and availability of surveillance cameras, respectively. Moreover, we propose a novel method for considering boundary information, resulting in reducing the percentage of predictions that pedestrians will be in an inappropriate area by half. In addition, our method can predict multiple pedestrians at once from the first-person perspective, leading to better prediction accuracy and the avoidance of collisions among pedestrians.

REFERENCES

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. (2016). Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*.

Amirian, J., Hayet, J.-B., and Pettré, J. (2019). Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Bhattacharyya, A., Hanselmann, M., Fritz, M., Schiele, B., and Straehle, C.-N. (2019). Conditional flow variational autoencoders for structured sequence prediction. arXiv.

Cai, Y., Dai, L., Wang, H., Chen, L., Li, Y., Sotelo, M. A., and Li, Z. (2022). Pedestrian motion trajectory prediction in intelligent driving from far shot first-person perspective video. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5298–5313.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *2020 European Conference on Computer Vision*.

Choi, C. and Dariush, B. (2019). Learning to infer relations for future trajectory forecast. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Deo, N. and Trivedi, M. M. (2020). Trajectory forecasts in unknown environments conditioned on grid-based plans. arXiv.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. (2018). Social gan: Socially acceptable trajectories with generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Huynh, M. and Alaghband, G. (2020). AOL: adaptive online learning for human trajectory prediction in dynamic video scenes. In *31st British Machine Vision Conference 2020*.

Katyal, K. D., Hager, G. D., and Huang, C.-M. (2020). Intent-aware pedestrian prediction for adaptive crowd navigation. In *2020 IEEE International Conference on Robotics and Automation*.

Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, H., and Savarese, S. (2019). Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H. S., and Chandraker, M. (2017). Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liang, J., Jiang, L., and Hauptmann, A. (2020). Simaug: Learning robust representations from simulation for trajectory prediction. In *2020 European Conference on Computer Vision*, Cham. Springer International Publishing.

Liang, J., Jiang, L., Niebles, J. C., Hauptmann, A. G., and Fei-Fei, L. (2019). Peeking into the future: Predicting future person activities and locations in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Makansi, O., Cicek, O., Buchicchio, K., and Brox, T. (2020). Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mangalam, K., Girase, H., Agarwal, S., Lee, K.-H., Adeli, E., Malik, J., and Gaidon, A. (2020). It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *2020 European Conference on Computer Vision*, Cham. Springer International Publishing.

Marchetti, F., Becattini, F., Seidenari, L., and Del Bimbo, A. (2020). Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Pang, B., Zhao, T., Xie, X., and Wu, Y. N. (2021). Trajectory prediction with latent belief energy-based model. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Poibrenski, A., Klusch, M., Vozniak, I., and Müller, C. (2020). *M2P3: Multimodal Multi-Pedestrian Path Prediction by Self-Driving Cars with Egocentric Vision*, page 190–197. Association for Computing Machinery, New York, NY, USA.

Qiu, J., Lo, F. P.-W., Gu, X., Sun, Y., Jiang, S., and Lo, B. (2021). Indoor future person localization from an egocentric wearable camera. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Rasouli, A., Kotseruba, I., Kunic, T., and Tsotsos, J. (2019). Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *2019 IEEE/CVF International Conference on Computer Vision*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*.

Rodin, I., Furnari, A., Mavroeidis, D., and Farinella, G. M. (2021). Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 211:103252.

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., and Arras, K. O. (2020). Human motion trajectory prediction: a survey. *The International Journal of Robotics Research*, 39(8):895–935.

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., and Savarese, S. (2019). Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Salzmann, T., Ivanovic, B., Chakravarty, P., and Pavone, M. (2020). Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *2020 European Conference on Computer Vision*, Cham. Springer International Publishing.

Seiskari, O., Rantalankila, P., Kannala, J., Ylilammi, J., Rahtu, E., and Solin, A. (2022). Hybvio: Pushing the limits of real-time visual-inertial odometry. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34.

Yagi, T., Mangalam, K., Yonetani, R., and Sato, Y. (2018). Future person localization in first-person videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yuan, Y., Chen, X., and Wang, J. (2020). Object-contextual representations for semantic segmentation. In *2020 European Conference on Computer Vision*.

Zhang, P., Ouyang, W., Zhang, P., Xue, J., and Zheng, N. (2019). Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhao, T., Xu, Y., Monfort, M., Choi, W., Baker, C., Zhao, Y., Wang, Y., and Wu, Y. N. (2019). Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.