M2P3: Multimodal Multi-Pedestrian Path Prediction by Self-Driving Cars With Egocentric Vision

Atanas Poibrenski
iMotion Germany GmbH
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
atanas.poibrenski@imotion.ai

Matthias Klusch
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
matthias.klusch@dfki.de

Igor Vozniak
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
igor.vozniak@dfki.de

Christian Müller
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
christian.mueller@dfki.de

ABSTRACT

Accurate prediction of the future position of pedestrians in traffic scenarios is required for safe navigation of an autonomous vehicle but remains a challenge. This concerns, in particular, the effective and efficient multimodal prediction of most likely trajectories of tracked pedestrians from egocentric view of self-driving car. In this paper, we present a novel solution, named M2P3, which combines a conditional variational autoencoder with recurrent neural network encoder-decoder architecture in order to predict a set of possible future locations of each pedestrian in a traffic scene. The M2P3 system uses a sequence of RGB images delivered through an internal vehicle-mounted camera for egocentric vision. It takes as an input only two modes, that are past trajectories and scales of pedestrians, and delivers as an output the three most likely paths for each tracked pedestrian. Experimental evaluation of the proposed architecture on the JAAD and ETH/UCY datasets reveal that the M2P3 system is significantly superior to selected state-of-the-art solutions.

KEYWORDS

Autonomous driving, multi-pedestrian path prediction

ACM Reference Format:
Atanas Poibrenski, Matthias Klusch, Igor Vozniak, and Christian Müller. 2020. M2P3: Multimodal Multi-Pedestrian Path Prediction by, Self-Driving Cars With Egocentric Vision. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341105.3373877

1 INTRODUCTION

Despite recent advances in autonomous driving, the achievement of pedestrian-safe navigation of autonomous vehicles (AVs) remains a challenge [47]. One prerequisite of collision-free navigation is an effective and efficient multi-pedestrian path prediction in traffic scenes by AVs. In fact, there is a plethora of solution approaches for this problem [65] to be employed in advanced driver assistance systems of AVs. Currently, these systems enable an AV to detect if a pedestrian is actually in the direction of travel, warn the control driver and even stop automatically. Other approaches would allow ADAS to predict whether the pedestrian is going to step on the street, or not [46].

The multimodality of multi-pedestrian path prediction in ego-view is a challenge and hard to handle by many deep learning (DL) models for many-to-one mappings. Given past trajectories of tracked pedestrians in a traffic scene, the distribution of future trajectories as outcomes has not a single but multiple modes. Each pedestrian has unique dynamics and individual goals to reach, and many different trajectory predictions are equally possible for the same traffic scene context with pedestrians. Conditional variational autoencoders (CVAE) for output representation learning and structured prediction may be applied to cope with this problem in principle [15]. A CVAE models the distribution of a high-dimensional output space as a generative model conditioned on input modes, which modulate the prior on lower dimensional, randomly sampled Gaussian latent variables that are then decoded into a set of probabilistic input reconstructions as outputs [35, 39, 55]. Though, the benefit of using a CVAE-based system for multimodal prediction of most likely pedestrian paths from egocentric vision of a self-driving car remains to be shown. It is not known for which set of input modes or factors of pedestrian dynamics, scene context and social context what kind of CVAE-based system architecture performs best for this purpose [47, 50].

To this end, we propose a novel CVAE-based system, named M2P3, for multimodal multi-pedestrian path prediction by self-driving cars in ego-view. It combines a conditional variational autoencoder as generative model with a recurrent neural network (RNN) encoder-decoder architecture in order to output a set of possible future paths of each pedestrian in a traffic scene tracked by a self-driving car with egocentric vision. The M2P3 system uses a RGB vehicle-mounted camera for egocentric vision and takes as input only two basic modes, that are past trajectories and scales of tracked pedestrians in traffic scene video. It k-means clusters the set of their trajectories predicted for 1 second into the future and outputs these k future

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3373877

pedestrian paths together with their probability of occurrence. Results of our comparative evaluation on the publicly available JAAD (joint attention for autonomous driving) ego-view video dataset reveal that the M2P3 system performance is significantly superior to selected state-of-the-art solutions.

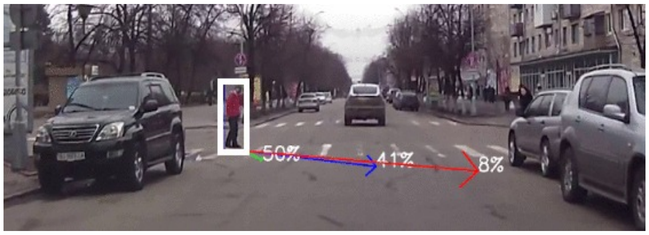

Figure 1: Example of M2P3 prediction of three most likely trajectories of tracked pedestrian in car ego-view video taken from the JAAD dataset.

The remainder of the paper is structured as follows. In Section 2, we briefly summarize related work and describe our novel solution M2P3 to the multi-pedestrian path prediction problem in Section 3. Results of our comparative experimental evaluation of M2P3 are discussed in Section 4 before we conclude in Section 5.

2 RELATED WORK

Egocentric vision. First-person video or egocentric vision is a subfield of computer vision which tries to analyze images or videos from a wearable camera (typically on a person's head) or from a mounted camera in car, looking forward. This is a challenging task due to the perspective view of the camera, the narrow field of view, as well as the introduced ego-motion. Most of the works in literature have focused on object detection [4, 29], activity recognition [13, 33, 36, 44],person identification [2, 11, 19, 60], activity forecasting [10, 16, 48], video summarization [28], gaze anticipation [32, 64], and grasp recognition [3, 6, 31, 52]. Recent work [41] also focuses on egocentric future localization but predicts the future location of the camera wearer and not the people around. Another example is the approach presented in [57], which uses a Siamese network to estimate future behaviors of basketball players in first-person videos. However, unlike our M2P3 approach, this method requires multiple cameras to reconstruct the scene.

Recent work that is more related to our M2P3 solution is presented in, for example, Bhattacharyya et al. [5]. The authors propose a Bayesian LSTM to predict the future locations of people by taking into account the car's ego-motion as well.

Yagi et al. [61] predict future locations of people observed in first-person videos by using the person's pose, past movement and ego-motion in a multi-stream convolution-deconvolution network. However, the method only predicts one possible future location, thus fails to capture multi-modality of the pedestrian motion.

Yao et al. [62] proposes a RNN encoder-decoder model that can predict future vehicle locations from ego-view in traffic scenarios. The approach employs scene optical flow as well as future ego-motion prediction but fails to model the probabilistic nature of the problem and takes no pedestrians into account.

Ma et al. [37] predicts the motion of heterogeneous traffic-agents from ego-view perspective using an LSTM-based realtime traffic prediction algorithm. They model the problem as a 4D graph and treat traffic agents as points and only take into account their past motions.

Trajectory prediction. The problem of human trajectory prediction has been researched extensively. Most of the works focus of static scenes and crowds. There are many classical approaches to the problem such as a Bayesian formulation [30, 53], Monte Carlo Simulation [8, 40, 49], Hidden Markov Models [14, 38], Kalman Filters [21], linear and non-linear Gaussian models [9, 45], Markov jump process [22]. These methods try to model objects based on their past movements but cannot work reliably in real-world traffic scenarios where uncertainty and multi-modality should be taken into account as well.

Other works explicitly model the interaction between pedestrians for collision avoidance. For example, in [43] the authors propose a linear trajectory avoidance model, and in [58] the social force model is utilized. These approaches are designed for homogenous interactions in crowds and rely on predetermined models of interaction.

In [1], a "Social LSTM" network is introduced, which predicts the future path of multiple people in a crowd by means of connected neighboring LSTMs in a social pooling layer. Recently, the authors of [17] propose to generate socially compliant set of trajectories by utilizing a GAN and training against a recurrent discriminator. However, their method is applied to a static crowd scene only.

Some recent work on pedestrian path prediction employ some variant of a recurrent neural network (RNN) and/or combine it with other deep learning models such as convolutional neural networks, generative adversarial networks (GANs), and variational autoencoders (VAE). For example, the DESIRE framework [27] consists of a CVAE-based RNN encoder-decoder architecture, which can output multiple path predictions to be refined further. However, the likelihood of each future path prediction per pedestrian is not estimated. The latter is achieved in M2P3 by means of k-means clustering to approximate the likelihood of future trajectories. Furthermore, according to our experiments the prior of DESIRE's CVAE appears too restrictive for modelling of multimodal trajectory distributions (cf. Table 2). More recent, the NEXT model [34] proposes a LSTM and focal attention-based approach to the prediction of trajectory and future activity of pedestrians. In particular, it combines visual features of person (appearance, pose), person-scene (segmentation of scene around person) and geometric person-object relations in a visual feature for separate trajectory generation with focal attention and action label prediction per pedestrian. However, in contrast to M2P3, NEXT does not address the above mentioned stochastic nature of the human trajectory prediction.

In [15], the use of conditional stochastic networks for multimodal prediction of object future motion trajectory in top-view with single frame as input from the drone Stanford dataset is investigated. However, our CVAE-based M2P3 system architecture and loss function are different, and implements a complete processing pipeline for self-driving car in ego-view.

3 M2P3 SOLUTION

As mentioned above, future prediction of pedestrian movement can be very ambiguous because given the same input state, multiple future states are possible. For example, a pedestrian heading towards a t-intersection, has an equal probability of going either left or right. Moreover, a model which simply learns a deterministic input/output mapping $f : \mathbb{X} \rightarrow \mathbb{Y}$ will under-represent the prediction space and possibly average out all possible outcomes, if a naive loss function is used. In order to tackle this problem, we adopt a generative model, especially a conditional variational auto-encoder (CVAE) based on gated recurrent neural networks for encoding and decoding, which generates a set of future pedestrian trajectories, hence allowing one-to-many input/output mappings.

3.1 Architecture

The architecture of the M2P3 approach is summarized in Figure 2. The pedestrian trajectory prediction problem is modeled with a generative model, a CVAE, where the posterior distribution $P(Y \mid X)$ is learned with the help of a latent variable Z [55]. Our model allows for conditional generation of pedestrian trajectories while taking into account the uncertainty of the future prediction.

The M2P3 gets as an input the trajectory and scale (represented as X) of each detected pedestrian in ego-view and predicts for each input for about two third of one second (n = 10 frames) the three most likely future trajectories $\hat{Y}$ for one second into the future (m = 15 frames). During training M2P3 is also provided with the ground truth of future pedestrian trajectories (given as Y). Its CVAE network learns to map the joint input H of RNN-based encodings of Y and X to latent Z with prior normal distribution in order to generate a model of $P(Y|X)$ which maximizes the probability of Y conditioned on input X. During testing, the processing of ground truth Y is removed such that only random samples Z from prior normal $\mathcal{N}(0, I)$ extended with encoded input X are used for prediction of Y with approximated posterior $P(Y|Z, X)$. For given number of N such predictions, the system eventually returns k most likely trajectories based on k-means clustering. The whole M2P3 processing pipeline allows to realize many-to-many mappings for multimodal multi-pedestrian path prediction in ego-view.

Training. The training architecture of M2P3 is shown in Figure 2. During training the time-dependent features of two basic input modes, that are pedestrian location l and scale s (see Sect. 3.2) of X, and the ground truth of future trajectory Y of the pedestrian are encoded through gated recurrent neural networks in H_X, respectively, H_Y. These encodings are concatenated in the joint input vector H for the variational module, which, in turn, learns estimating the mean μ_H and co-variance Σ_H of normal distribution $\mathcal{N}(\mu_H, \Sigma_H)$, mapping the joint input H to latent Z with conditional normal prior $P(Z|X) \sim \mathcal{N}(0, I)$ as reference for sampling. A random sample of Z from normal distribution together with condition H_X is then fed into the following RNN decoder. The latter decodes this extended sample into a predicted future trajectory with approximated conditional normal posterior distribution $P(Y|Z, X)$.

Each of the encoding of inputs X and Y into H_Y and H_X is done by a RNN encoder using the following recurrent (GRU) computation:

$$\begin{aligned} \mathbf{z}_t &= \sigma(\mathbf{W}_z \cdot \mathbf{x}_t + \mathbf{U}_z \cdot \mathbf{h}_{t-1} + \mathbf{b}_z), \\ \mathbf{r}_t &= \sigma(\mathbf{W}_r \cdot \mathbf{x}_t + \mathbf{U}_r \cdot \mathbf{h}_{t-1} + \mathbf{b}_r), \\ \mathbf{h}_t &= (1 - \mathbf{z}_t)\mathbf{h}_{t-1} + \mathbf{z}_t\sigma(\mathbf{W}_h \cdot \mathbf{x}_t + \mathbf{U}_h(\mathbf{r}_t\mathbf{h}_{t-1}) + \mathbf{b}_h) \end{aligned} \tag{1}$$

where $\mathbf{W}$, $\mathbf{U}$ and $\mathbf{b}$ are learnable weights, $\mathbf{z}$ and $\mathbf{r}$ are update and reset gates; $\mathbf{x}$ and $\mathbf{h}$ are input and output vectors accordingly. σ is a nonlinear function such as tanh. Initially, for $\mathbf{t} = 0$, the output vector is $\mathbf{h}_0 = 0$.

The outputs H_Y, H_X of both encoders are then concatenated into a joint input vector H. This vector is fed into two fully-connected layers for mean μ_H and co-variance Σ_H, which are learned to model the latent Z distribution $Q(Z|H)$ as normal distribution $\mathcal{N}(\mu_H, \Sigma_H)$ with $Z = \mu_H + \Sigma_H \odot \epsilon$ and $\epsilon \sim \mathcal{N}(0, I)$. In other words, it learns to map the joint input H to latent Z with normal distribution $\mathcal{N}(0, I)$ as reference for sampling; $P(Z|X)$ is $\mathcal{N}(0, I)$, because we assume Z is sampled independently of X at test time. This processing part of the M2P3-CVAE network during learning requires the minimization of the Kullback-Leibler divergence (D_{KL}) between the estimated distribution $Q(Z|H)$ and the reference distribution $\mathcal{N}(0, I)$, i.e. $D_{KL}(\mathcal{N}(\mu_H, \Sigma_H)||\mathcal{N}(0, I))$.

In order to allow for backpropagation of errors through a layer that samples Z from $Q(Z|H)$, which is a non-continuous operation without gradient, the standard reparameterization trick to move the sampling to an input layer as introduced in [24] is applied. That is, sampling from $\mathcal{N}(\mu_H, \Sigma_H)$ is done by first randomly sampling $\epsilon \sim \mathcal{N}(0, I)$ and then computing Z with these parameters $(\epsilon, \mu_H, \Sigma_H)$ as mentioned above.

Eventually, the RNN decoder gets a sample of Z extended with condition H_X, performs the recurrent operation (1) on it, and feeds the result into a final dense layer that produces the future trajectory prediction $\hat{Y}$. This processing part of the M2P3-CVAE network during learning requires the minimization of the error between ground truth future trajectory Y and its prediction $\hat{Y}$ according to the L2 loss (Euclidean distance) $\left\|Y - \hat{Y}\right\|^2$.

The whole M2P3-CVAE network architecture is trained with stochastic gradient descent method to minimize the total loss L defined as

$$L = \left\|Y - \hat{Y}\right\|^2 + \mathrm{D_{KL}}(\mathcal{N}(\mu_H, \Sigma_H)||\mathcal{N}(0, I)) \tag{2}$$

That is, the latent distribution Q is learned by the M2P3-CVAE network such that it gives a higher probability to Z with which it is more likely to produce predictions $\hat{Y}$ that are close to ground truth Y in the context of X.

Testing. The M2P3 test architecture is shown in Figure 3. At test time, the ground truth of future trajectories Y is not available such that the respective part of the encoding pathway in the M2P3 system is not used (see Figure 3). Besides, we can now sample from distribution $P(Y|X)$ by sampling $Z \sim \mathcal{N}(0, I)$ In fact, the RNN decoder of M2P3 only receives the RNN-encoded condition H_X together with a random sample Z drawn from the prior distribution $\mathcal{N}(0, I)$. This enables probabilistic inference allowing to handle multimodality in the prediction space.

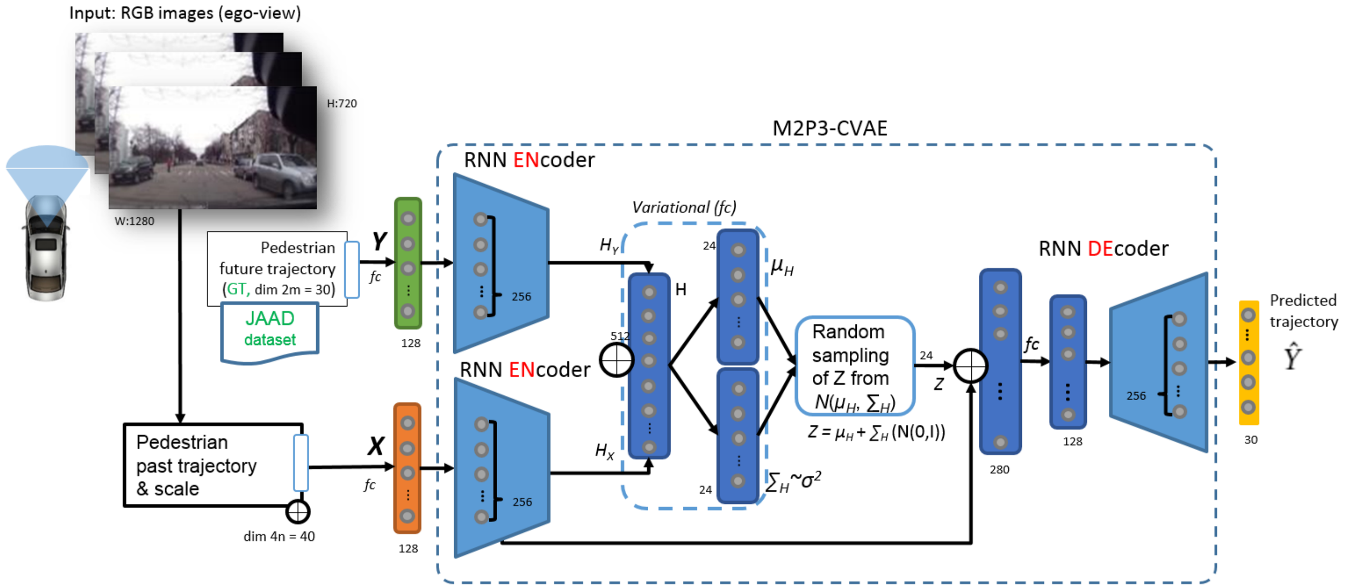

Figure 2: M2P3 system architecture (training) overview. For each tracked pedestrian, the system processes the ground-truth future trajectory (Y) from JAAD dataset, and observed past trajectory (X) by means of encoding, mapping of joint input H to latent Z with normal prior as reference for sampling, and decoding of random sample Z with X into prediction of future pedestrian trajectory $\hat{Y}$ as output.

For each input X the test network is run with N = 1000 random samples Z, thereby generating N possible trajectories of the considered pedestrian, which are then clustered into k = 3 clusters using k-means. Since we do not have explicit access to the posterior trajectory distribution, we choose a large number for N, which allows the future trajectory distribution to be closely approximated. The value of k is chosen arbitrary, such that the output trajectories are not under or over-clustered. In particular, each of the generated trajectories is assigned to the closest cluster based on 2D Euclidean distance. The number of assigned trajectories in each cluster is divided by the total number N of generated trajectories to obtain a probability distribution over the clusters (see Figure 1). In concrete terms, given the set of output trajectories from the model $Y = \{\mathrm{Y}_1, \mathrm{Y}_2, ..., \mathrm{Y}_N\}$ and the set $\mathrm{S} = \{\mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3\}$ of clusters, M2P3 assigns each output trajectory Y_p, $1 \leq p \leq \mathrm{N}$, to exactly one cluster S_i, $1 \leq i \leq 3$, whose mean m_i has the least squared Euclidean distance:

$$\mathrm{S}_i = \{\mathrm{Y}_p : \left\|\mathrm{Y}_p - \mathrm{m}_i\right\|^2 \leq \left\|\mathrm{Y}_p - \mathrm{m}_j\right\|^2, 1 \leq j \leq 3, i \neq j\} \quad (3)$$

This is followed by the calculation of (k=3) cluster probabilities as $P(\mathrm{S}_i) = |\mathrm{S}_i|/|\mathrm{Y}|$, where $|\mathrm{S}_i|$ and $|\mathrm{Y}|$ denote the cardinality of respective sets. These cluster probabilities are then displayed by the M2P3 system as probabilities of occurrence of predicted future pedestrian positions.

Implementation. Our M2P3 implementation bases on Keras [7] with Tensorflow as backend. For pedestrian tracking, M2P3 can utilize DeepSORT [59] with underlying mask R-CNN [18] for pedestrian detection. In the implemented M2P3 system, all CVAE input data first passes through a fully-connected embedding layer of size 128 before being fed into an encoder. The hidden size of all encoders and decoder layers is set to 256. The latent dimension of the fully-connected (fc) layer in the CVAE is set to 24. The two latent fc layers are concatenated before deriving latent distribution, that is to match the unknown latent distribution to a known, prior distribution. In order to simplify the training process, in contrast to Long-Short-Term-Memory (LSTM) networks, Gated Recurrent Units (GRU) have been adopted for the RNN-encoders/-decoder.

3.2 Pedestrian Trajectory and Scale

One obvious clue about future pedestrian motions is their motion in the past. Thus, M2P3 also tracks each pedestrian's 2D image location (x,y coordinates) for **n** frames. For each detected pedestrian in the scene, M2P3 collects the following feature vector:

$\mathrm{X}_\mathrm{l} = \{\mathrm{x}_{\mathrm{T-n}}, \mathrm{y}_{\mathrm{T-n}}, \mathrm{x}_{\mathrm{T-(n-1)}}, \mathrm{y}_{\mathrm{T-(n-1)}}, ..., \mathrm{x}_\mathrm{T}, \mathrm{y}_\mathrm{T}\}$,

where T is the current time frame. 2D image distances correspond to different physical distances depending on where the person is situated in the frame. Therefore, M2P3 learns the width and the height (scale) of the pedestrian in order to take the perspective effect of the ego camera into account. In particular, it records the width w and height h in pixels of each pedestrian for the past **n** frames into the following vector:

$\mathrm{X}_\mathrm{s} = \{\mathrm{w}_{\mathrm{T-n}}, \mathrm{h}_{\mathrm{T-n}}, \mathrm{w}_{\mathrm{T-(n-1)}}, \mathrm{h}_{\mathrm{T-(n-1)}}, ..., \mathrm{w}_\mathrm{T}, \mathrm{h}_\mathrm{T}\}$

The final input $\mathrm{X}_{\mathrm{l,s}}$ to the underlying M2P3 model (cf. Sect. 3.1) then is: $\mathrm{X}_{\mathrm{l,s}} = \mathrm{X}_\mathrm{l} \oplus \mathrm{X}_\mathrm{s}$, where $\oplus$ denotes the concatenation operator. This input is normalized in the range [0,1] relative to the image resolution. The output Y of the M2P3 model is modeled as the 2D displacement from the last observed frame T:

$\mathrm{Y} = \{\mathrm{x}_{\mathrm{T+1}} - \mathrm{x}_\mathrm{T}, \mathrm{y}_{\mathrm{T+1}} - \mathrm{y}_\mathrm{T}, ..., \mathrm{x}_{\mathrm{T+m}} - \mathrm{x}_\mathrm{T}, \mathrm{y}_{\mathrm{T+m}} - \mathrm{y}_\mathrm{T}\}$,

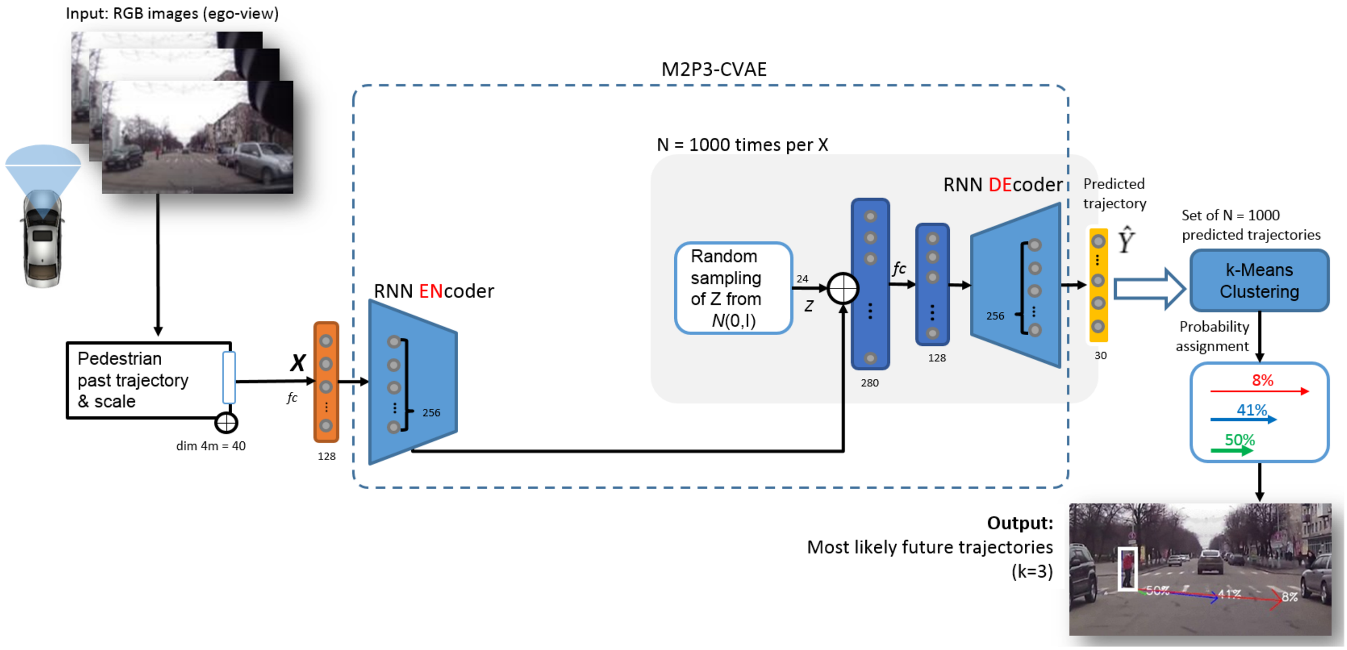

Figure 3: M2P3 architecture (testing) overview. The input only consists of X for observed pedestrian trajectory and scale, which RNN-based encoding combined with random sample of Z from normal distribution is decoded into trajectory prediction $\hat{Y}$ as output.

where **m** is the number of frames in the future. By using a displacement vector rather than absolute coordinates, the M2P3 model is able to learn how a pedestrian moves in the future relative to his starting position. This helps with the generalization of the model to new scenes with different resolution and positions of the pedestrians.

4 EXPERIMENTS

For comparative performance evaluation of our M2P3 system, we conducted experiments based on the publicly available datasets JAAD, ETH, and UCY against selected state-of-the-art multi-pedestrian path predictors as baselines.

4.1 JAAD

Dataset. For our first comparative performance evaluation experiments, we use the publicly available JAAD (Joint Attention for Autonomous Driving) dataset [25]. This dataset contains an annotated collection of short video clips, capturing typical urban traffic scenarios in various weather conditions. The clips are taken from a single RGB camera, mounted behind the windshield of a moving car. All pedestrians are manually annotated with bounding boxes and unique tracking identifier. The resolution of all videos is set to a constant value of 1280 × 720. The frame rate is also re-scaled to a constant value of n = 15 frames per second. All pedestrians which are either too far away from the car (less than 50 pixels in size), or occluded, or tracked for less than 25 frames, are ignored.

Implementation. The JAAD dataset is split into training (videos 0-250) and testing (videos 251-346) as done in [12, 56]. The ratio between training and validation videos is 80% to 20% for fine-tuning the hyper-parameters of the implemented M2P3 model. After hyper-parameters are fixed, we train the M2P3 on the full training set of JAAD (videos 0-250).

For all experiments, ground truth bounding boxes provided by the JAAD dataset are used for extracting past trajectory and scale of pedestrian. The numbers of past and future frames are set to $n = 15, m = 10$. The ADAM [23] optimizer is used with learning rate of 1e-4 and trained the M2P3 for 6000 epochs. The model has 948,914 trainable parameters in total. The training takes approximately 2 hours on desktop machine with NVIDIA GTX 1080ti GPU and Intel i7-7800X CPU. The average inference time is 29ms per pedestrian.

Baselines and Metrics. For the comparative performance evaluation, we selected the following five state-of-the-art solution models as baselines.

(1) **CV** (Constant Velocity) model. The CV model as in [54] assumes that the pedestrian maintains constant velocity through time. The horizontal and vertical components of the velocity at time t are denoted as v_t^x and v_t^y and defined as $v_t^x = \frac{x_t - x_{t-n}}{n}$ and $v_t^y = \frac{y_t - y_{t-n}}{n}$ for the past n observed frames. Therefore, the future position of a pedestrian is defined as $\tilde{x}_{t+m} = x_t + v_t^x \cdot m$ and $\tilde{y}_{t+m} = y_t + v_t^y \cdot m$ in the next m frames.

(2) **CA** (Constant Acceleration) model. The implemented CA model is the same as the CV model above but the acceleration of a pedestrian is assumed to be constant.
(3) **RNN** Encoder-decoder model in [62] is the same as the M2P3 model but without the CVAE module.
(4) **MSCD** model. The MSCD model proposed in [61] uses a multi-stream convolution-deconvolution framework.
(5) **DTP** model. The DTP model proposed in [56] utilizes the optical flow of the pedestrians and a residual network.

For reasons of comparability of the results, we applied the same evaluation scheme as in [56]. The M2P3 model observes **n** = 10 frames (2/3 of a second) and predicts **m** = 15 frames (1 second) into the future. The following performance evaluation metrics for pedestrian path prediction are computed for all systems on the test set of the JAAD dataset:

- The mean squared error (MSE) in pixels (1280x720 resolution)is defined as $\frac{1}{N}\frac{1}{m}\sum_{i=1}^{N}\sum_{t=1}^{m}(\hat{\mathrm{Y}^{\mathrm{i}}_{\mathrm{t}}}-\mathrm{Y}^{\mathrm{i}}_{\mathrm{t}})^2$, where Ŷ denotes the prediction, Y the ground truth, and N the number of test sequences.
- The displacement error (DE) in pixels (1280x720 resolution) at last time step **m** = 15 (1 second) is defined as $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{Y}^i_m - Y^i_m\right\|_2$.

Results and Analysis. The results of the comparative performance evaluation over the JAAD dataset are given in Table 1.

Method	MSE	DE
CA[54]	1426	52.8
CV	1148	47.5
RNN [62]	983	49.1
MSCD [61]	881 ± 44	41.3 ± 1.2
DTP [56]	610 ± 21	34.6 ± 0.5
M2P3 (1 sample)	584 ± 5	35.9 ± 0.15
M2P3 (1000 samples, k=3 clusters)	**483 ± 2**	**29.02 ± 0.06**

Table 1: Experimental results for M2P3 and baselines over the JAAD dataset

The results reveal that with just a single random sample (prediction) our CVAE-RNN based model M2P3 is already able to reach a performance comparable to that of the selected state-of-the-art baselines. When the number of samples is increased to 1000, clustered into 3 clusters, where the cluster closest to the ground truth is chosen, our model notably even outperforms all selected baselines. This confirms the multi-modal nature of the prediction problem, which is not captured by the selected alternative methods. Having access to the ground truth of pedestrian path prediction in the real world of autonomous driving is, of course, not possible, hence one cannot simply pick the best prediction. This case is handled by the M2P3 by means of clustering of and probability assignment to the predictions in the output set. These clusters can then be considered one by one in a decreasing probability fashion by an AV navigation algorithm. Besides, the M2P3 also implicitly learns both pedestrian and ego motion instead of ego motion-free trajectories for learning individual motion patterns of pedestrians.

4.2 ETH/UCY

Dataset. For our second comparative performance evaluation, we used two prominent, publicly available datasets for trajectory prediction: ETH [57] and UCY [26]. Both datasets are converted to world coordinates (meters) and pedestrian positions are obtained every 0.4 seconds (1 timeframe). The data is split into 5 sets (ETH - 2, UCY - 3) and we follow the standard leave-one-scene-out data split as in [17] for evaluation, such that training is performed on 4 sets and test on the remaining one. Past trajectories are observed for 8 timesteps (3.2 seconds) and predicted for the next 12 timesteps (4.8 seconds).

Implementation. Here, we experimented with a more powerful prior, that is the Mixture-of-Gaussians (MoG), which can capture more modes of the trajectory distribution compared to just a unit Gaussian. The loss function L for M2P3-MoG training is as follows:

$$L = \left\|Y - \hat{Y}\right\|^2 + \mathrm{D_{KL}}(\mathrm{q(z, c|X, Y)}||p(\mathrm{z, c|X})) \qquad (4)$$

This essentially means that the variational encoder of the M2P3 now learns a posterior distribution q(z,c|X,Y), where the latent embedding **z** is regularized by the prior p(z,c|X) to lie on Mixture-of-Gaussians manifold, where $\mathbf{z} \sim \mathcal{N}(\mu_c,\ \sigma_c^2\mathrm{I})$ and $\mathbf{c} \sim$ Category(π) such that K is a predefined number of components of the mixture and π=[π_1, π_2,...,π_K] is the prior probability of the Gaussian mixture components. More details for the derivation of the loss function can be seen in [20], where a variational autoencoder with MoG was originally used for the task of clustering.
Since the ETH/UCY dataset is captured from a fixed top-down view, we do not use the person's scale anymore but the normalized past trajectory in meters. For a stable training, the model gets first pre-trained based on just the first term of the loss in (4) for a few epochs. After that a gaussian mixture from the latent space (Z) of the model is initialized for continued training with the full loss for 100 epochs and the ADAM optimizer with a learning rate of 1e-5.

Baselines and Metrics. For the evaluation, 20 predictions are generated for each observed trajectory and the closest one to the ground truth is chosen. This allows us to test the multi-modality and diversity of the predictions. We compare our model to the following state-of-the-art baselines:

(1) **Social GAN** [17] uses a recurrent sequence-to-sequence model with a novel social pooling mechanism and a generative adversarial network.
(2) **Sophie** [51] uses a generative adversarial network to generate realistic trajectory by utilizing social and physical scene constraints.
(3) **NEXT** [34] uses a LSTM encoder-decoder architecture to predict persons' movements and utilizes rich visual features about human behavioral information and interaction with their surroundings.
(4) **DESIRE** [27] combines a RNN encoder-decoder with a CVAE and uses the person's past trajectory and scene context to predict the future trajectory.

ADE/FDE in meters (20 samples)						
Method	ETH	HOTEL	UNIV	ZARA1	ZARA2	Average
Social GAN [17]	0.81/1.52	0.72/1.61	0.60/1.26	0.34/0.69	0.42/0.84	0.58/1.18
Sophie [51]	0.70/1.43	0.76/1.67	0.54/1.24	**0.30/0.63**	0.38/0.78	0.54/1.15
NEXT [34]	0.73/1.65	**0.30/0.59**	0.60/1.27	0.38/0.81	0.31/0.60	0.46/1.00
DESIRE [27]	0.93/1.94	0.52/1.03	0.59/1.27	0.41/0.86	0.33/0.72	0.53/1.11
SGN LSTM [63]	0.75/1.63	0.63/1.01	**0.48/1.08**	**0.30**/0.65	0.26/0.57	0.48/0.99
FSGAN [42]	0.68/1.16	0.43/0.89	0.54/1.14	0.35/0.71	0.32/0.67	0.46/0.91
M2P3 (Ours)	1.04/2.16	0.54/1.13	0.64/1.34	0.45/0.95	0.37/0.79	0.60/1.27
M2P3 MoG (Ours)	**0.57/1.01**	0.40/0.87	0.61/1.31	0.33/0.70	**0.21**/**0.42**	**0.42/0.86**

Table 2: Experimental results for M2P3 and baselines over the ETH/UCY dataset. For each method the best out of 20 predictions (samples) is chosen.

(5) **SGN LSTM** [63] is a stochastic trajectory predictor which uses LSTM and directed social graph which is dynamically constructed on timely location and speed direction.
(6) **FSGAN** [42] extends Social GAN [17] by incorporating adversarial loss in the trajectory prediction task.

For comparison, we adopt the error metrics from prior work [1, 27]:

(1) *Average Displacement Error* (ADE) is the average L2 distance between the prediction and the ground truth over all time steps.
(2) *Final Displacement Error* (FDE) is the L2 distance between the prediction and the ground truth at the last time step (in our experiments: 4.8 seconds).

Results and Analysis. The results for the ETH/UCY dataset are summarized in Table 2. Our M2P3 model with a unit Gaussian prior performs the worst as it is unable to fully capture all of the modes of trajectory distribution. Its predictions are simply forced around the mean of this single Gaussian. However, by exchanging the prior with a Mixture-of-Gaussians one, the M2P3-MoG was able to successfully capture multiple modes of the data. Even though the M2P3-MoG uses only the past trajectory as an input, it achieved the lowest prediction error in our experiments. This suggests that a generative model with a diverse prior is crucial for achieving state-of-the-art results on this particular dataset. A disadvantage of the MoG prior is that one needs to manually choose the amount of mixture components (in this case 5) but that can be addressed by hyper-parameter tuning on a validation set.

5 CONCLUSIONS

In this paper, we presented a novel solution M2P3 for the egocentric multi-modal multi-pedestrian path prediction problem. M2P3 combines a conditional variational autoencoder with a recurrent neural network encoder-decoder architecture. It uses a RGB vehicle-mounted camera for egocentric vision, takes two inputs by computing past trajectories and scales of tracked pedestrians in the field of car perception with egocentric vision and then outputs diverse trajectories together with their probability of occurrence. Results of comparative experimental evaluation on the JAAD dataset showed that the M2P3 model can outperform selected state-of-the-art solutions. Furthermore, the M2P3 with a simple change of the prior to a Mixture-of-Gaussians already showed comparable performance to that of more complex state-of-the-art path predictors over the prominent ETH/UCY dataset. Ongoing work is concerned with the separation of car ego motion from pedestrian motion, and the ablative investigation of integrating additional factors of pedestrian intention estimation and interaction. Additionally, increasing the diversity of the output as well as incorporating even more sophisticated prior, will be further investigated.

Acknowledgement. This research was supported by iMotion Germany GmbH and the German Federal Ministry for Education and Research (BMB+F) in the project REACT.

REFERENCES

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. 2016. Social LSTM: Human Trajectory Prediction in Crowded Spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 961–971. https://doi.org/10.1109/CVPR.2016.110
[2] Shervin Ardeshir and Ali Borji. 2016. Ego2Top: Matching Viewers in Egocentric and Top-view Videos. In *ECCV*.
[3] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. 2015. Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), 1949–1957.
[4] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. 2017. First-Person Action-Object Detection with EgoNet. *ArXiv* abs/1603.04908 (2017).
[5] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. 2018. Long-Term Onboard Prediction of People in Traffic Scenes Under Uncertainty. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 4194–4202.
[6] Minjie Cai, Kris Makoto Kitani, and Yoichi Sato. 2015. A scalable approach for understanding the visual structures of hand grasps. *2015 IEEE International Conference on Robotics and Automation (ICRA)* (2015), 1360–1366.
[7] François Chollet et al. 2015. Keras. https://keras.io.
[8] S. Danielsson, L. Petersson, and A. Eidehall. 2007. Monte Carlo based Threat Assessment: Analysis and Improvements. In *2007 IEEE Intelligent Vehicles Symposium*. 233–238. https://doi.org/10.1109/IVS.2007.4290120
[9] D. Ellis, E. Sommerlade, and I. Reid. 2009. Modelling pedestrian trajectory patterns with Gaussian processes. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. 1229–1234. https://doi.org/10.1109/ICCVW.2009.5457470
[10] Chenyou Fan, Jangwon Lee, and Michael S. Ryoo. 2017. Forecasting Hand and Object Locations in Future Frames. *CoRR* abs/1705.07328 (2017). arXiv:1705.07328 http://arxiv.org/abs/1705.07328
[11] C. Fan, J. Lee, M. Xu, K. K. Singh, Y. J. Lee, D. J. Crandall, and M. S. Ryoo. 2017. Identifying First-Person Camera Wearers in Third-Person Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4734–4742. https://doi.org/10.1109/CVPR.2017.503
[12] Zhijie Fang and Antonio M. López. 2018. Is the Pedestrian going to Cross? Answering by 2D Pose Estimation. *2018 IEEE Intelligent Vehicles Symposium (IV)* (2018), 1271–1276.
[13] Alireza Fathi, Ali Farhadi, and James M. Rehg. 2011. Understanding egocentric activities. *2011 International Conference on Computer Vision* (2011), 407–414.
[14] J. Firl, H. StÃijbing, S. A. Huss, and C. Stiller. 2012. Predictive maneuver evaluation for enhancement of Car-to-X mobility data. In *2012 IEEE Intelligent Vehicles Symposium*. 558–564. https://doi.org/10.1109/IVS.2012.6232217

[15] Katerina Fragkiadaki, Jonathan Huang, Alex Alemi, Sudheendra Vijayanarasimhan, Susanna Ricco, and Rahul Sukthankar. 2017. Motion prediction under multimodality with conditional stochastic networks. *arXiv preprint arXiv:1705.02082* (2017).

[16] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. 2017. Next-Active-Object prediction from Egocentric Videos. *ArXiv* abs/1904.05250 (2017).

[17] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. 2018. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. *CoRR* abs/1803.10892 (2018). http://dblp.uni-trier.de/db/journals/corr/corr1803.html#abs-1803-10892

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[19] Yedid Hoshen and Shmuel Peleg. 2014. Egocentric Video Biometrics. *CoRR* abs/1411.7591 (2014). arXiv:1411.7591 http://arxiv.org/abs/1411.7591

[20] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. 2016. Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering. In *IJCAI*.

[21] Rudolf E. Kálmán. 1960. A New Approach to Linear Filtering and Prediction.

[22] Vasiliy Karasev, Alper Ayvaci, Bernd Heisele, and Stefano Soatto. 2016. Intent-aware long-term prediction of pedestrian motion. *2016 IEEE International Conference on Robotics and Automation (ICRA)* (2016), 2543–2549.

[23] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2015).

[24] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2014).

[25] Iuliia Kotseruba, Amir Rasouli, and John K. Tsotsos. 2016. Joint Attention in Autonomous Driving (JAAD). *arXiv e-prints*, Article arXiv:1609.04741 (Sep 2016), arXiv:1609.04741 pages. arXiv:cs.RO/1609.04741

[26] Laura Leal-TaixÃľ, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. 2014. Learning an Image-Based Motion Context for Multiple People Tracking. https://doi.org/10.1109/CVPR.2014.453

[27] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher Bongsoo Choy, Philip H. S. Torr, and Manmohan Krishna Chandraker. 2017. DESIRE: Distant Future Prediction in Dynamic Scenes with Interacting Agents. *CoRR* abs/1704.04394 (2017). arXiv:1704.04394 http://arxiv.org/abs/1704.04394

[28] Yong Jin Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 1346–1353.

[29] Yong Jin Lee and Kristen Grauman. 2014. Predicting Important Objects for Egocentric Video Summarization. *International Journal of Computer Vision* 114 (2014), 38–55.

[30] S. LefÃĺvre, C. Laugier, and J. IbaÃśez-GuzmÃąn. 2011. Exploiting map information for driver intention estimation at road intersections. In *2011 IEEE Intelligent Vehicles Symposium (IV)*. 583–588. https://doi.org/10.1109/IVS.2011.5940452

[31] Cheng Yen Li and Kris M. Kitani. 2013. Pixel-Level Hand Detection in Ego-centric Videos. *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013), 3570–3577.

[32] Yin Li, Alireza Fathi, and James M. Rehg. 2013. Learning to Predict Gaze in Egocentric Video. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV ’13)*. IEEE Computer Society, Washington, DC, USA, 3216–3223. https://doi.org/10.1109/ICCV.2013.399

[33] Yin Li, Zhefan Ye, and James M. Rehg. 2015. Delving into egocentric actions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 287–295.

[34] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G. Hauptmann, and Li Fei-Fei. 2019. Peeking into the Future: Predicting Future Person Activities and Locations in Videos. *CoRR* abs/1902.03748 (2019). arXiv:1902.03748 http://arxiv.org/abs/1902.03748

[35] Manuel Lopez-Martin, Belen Carro, Antonio Sanchez-Esguevillas, and Jaime Lloret. 2017. Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot. *Sensors* 17, 9 (2017), 1967.

[36] Minghuang Ma, Haoqi Fan, and Kris Makoto Kitani. 2016. Going Deeper into First-Person Activity Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1894–1903.

[37] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. 2019. TrafficPredict: Trajectory Prediction for Heterogeneous Traffic-Agents. *ArXiv* abs/1811.02146 (2019).

[38] Dimitrios Makris and Tim J. Ellis. 2002. Spatial and Probabilistic Modelling of Pedestrian Behaviour. In *BMVC*.

[39] Ashish Mishra, Shiva Krishna Reddy, Anurag Mittal, and Hema A Murthy. 2018. A generative model for zero shot learning using conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2188–2196.

[40] Sang Min Oh, James M. Rehg, Tucker R. Balch, and Frank Dellaert. 2007. Learning and Inferring Motion Patterns using Parametric Segmental Switching Linear Dynamic Systems. *International Journal of Computer Vision* 77 (2007), 103–124.

[41] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. 2016. Egocentric Future Localization. (June 2016).

[42] Alexandre Alahi Parth Kothari. 2019. Human Trajectory Prediction using Adversarial Loss.

[43] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. 2009. You’ll never walk alone: Modeling social behavior for multi-target tracking. *2009 IEEE 12th International Conference on Computer Vision* (2009), 261–268.

[44] Hamed Pirsiavash and Deva Ramanan. 2012. Detecting activities of daily living in first-person camera views. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 2847–2854.

[45] Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[46] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. 2017. Agreeing to cross: How drivers and pedestrians communicate. *2017 IEEE Intelligent Vehicles Symposium (IV)* (2017), 264–269.

[47] A. Rasouli and J. K. Tsotsos. 2019. Autonomous Vehicles That Interact With Pedestrians: A Survey of Theory and Practice. *IEEE Transactions on Intelligent Transportation Systems* (2019), 1–19. https://doi.org/10.1109/TITS.2019.2901817

[48] Nicholas Rhinehart and Kris Makoto Kitani. 2017. First-Person Activity Forecasting with Online Inverse Reinforcement Learning. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 3716–3725.

[49] A. V. I. Rosti and M. J. F. Gales. 2004. Rao-Blackwellised Gibbs sampling for switching linear dynamical systems. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. I–809. https://doi.org/10.1109/ICASSP.2004.1326109

[50] A. Rudenko and et al. 2019. Human Motion Trajectory Prediction: A Survey. In *arXiv preprint arXiv:1905.06113*.

[51] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, and Silvio Savarese. 2018. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints. *CoRR* abs/1806.01482 (2018). arXiv:1806.01482 http://arxiv.org/abs/1806.01482

[52] A. Saran, D. Teney, and K. M. Kitani. 2015. Hand parsing for fine-grained recognition of human grasps in monocular images. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5052–5058. https://doi.org/10.1109/IROS.2015.7354088

[53] Nicolas Schneider and Dariu M. Gavrila. 2013. Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study. In *Pattern Recognition*, Joachim Weickert, Matthias Hein, and Bernt Schiele (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 174–183.

[54] Christoph Schöller, Vincent Aravantinos, Florian Lay, and Alois Knoll. 2019. The Simpler the Better: Constant Velocity for Pedestrian Motion Prediction. *ArXiv* abs/1903.07933 (2019).

[55] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*. 3483–3491.

[56] Olly Styles, Arun Ross, and Víctor Rojo Sánchez. 2019. Forecasting Pedestrian Trajectory with Machine-Annotated Training Data. *ArXiv* abs/1905.03681 (2019).

[57] Shan Su, Jung Pyo Hong, Jianbo Shi, and Hyun Soo Park. 2017. Predicting Behaviors of Basketball Players from First Person Videos. 1206–1215. https://doi.org/10.1109/CVPR.2017.133

[58] Jur P. van den Berg, Stephen J. Guy, Ming C. Lin, and Dinesh Manocha. 2009. Reciprocal n-Body Collision Avoidance. In *ISRR*.

[59] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple Online and Realtime Tracking with a Deep Association Metric. *CoRR* abs/1703.07402 (2017). arXiv:1703.07402 http://arxiv.org/abs/1703.07402

[60] Mingze Xu, Chenyou Fan, Yuchen Wang, Michael S. Ryoo, and David J. Crandall. 2018. Joint Person Segmentation and Identification in Synchronized First- and Third-Person Videos. In *ECCV*.

[61] Takuma Yagi, Karttikeya Mangalam, Ryo Yonetani, and Yoichi Sato. 2018. Future Person Localization in First-Person Videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 7593–7602.

[62] Yu Yao, Mingze Xu, Chiho Choi, David J. Crandall, Ella M. Atkins, and Behzad Dariush. 2018. Egocentric Vision-based Future Vehicle Localization for Intelligent Driving Assistance Systems. *CoRR* abs/1809.07408 (2018). arXiv:1809.07408 http://arxiv.org/abs/1809.07408

[63] Lidan Zhang, Qi She, and Ping Guo. 2019. Stochastic trajectory prediction with social graph network. *CoRR* abs/1907.10233 (2019). arXiv:1907.10233 http://arxiv.org/abs/1907.10233

[64] M. Zhang, K. T. Ma, J. H. Lim, Q. Zhao, and J. Feng. 2017. Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3539–3548. https://doi.org/10.1109/CVPR.2017.377

[65] Yue Zhang, Yonggang Qi, Jun Liu, and Yanyan Wang. 2018. Decade of Vision-Based Pedestrian Detection for Self-Driving: An Experimental Survey and Evaluation, In SAE Technical Paper. https://doi.org/10.4271/2018-01-1603