Multimodal Future Localization and Emergence Prediction for Objects in Egocentric View with a Reachability Prior

Osama Makansi[1] Özgün Çiçek[1] Kevin Buchicchio[2] Thomas Brox[1]
[1]University of Freiburg
[2]IMRA-EUROPE
makansio,cicek,brox@cs.uni-freiburg.de buchicchio@imra-europe.com

Abstract

In this paper, we investigate the problem of anticipating future dynamics, particularly the future location of other vehicles and pedestrians, in the view of a moving vehicle. We approach two fundamental challenges: (1) the partial visibility due to the egocentric view with a single RGB camera and considerable field-of-view change due to the egomotion of the vehicle; (2) the multimodality of the distribution of future states. In contrast to many previous works, we do not assume structural knowledge from maps. We rather estimate a reachability prior for certain classes of objects from the semantic map of the present image and propagate it into the future using the planned egomotion. Experiments show that the reachability prior combined with multi-hypotheses learning improves multimodal prediction of the future location of tracked objects and, for the first time, the emergence of new objects. We also demonstrate promising zero-shot transfer to unseen datasets.

1. Introduction

Figure 1 shows the view of a driver approaching pedestrians who are crossing the street. To safely control the car, the driver must anticipate where these pedestrians will be in the next few seconds. Will the last pedestrian (in blue) have completely crossed the street when I arrive or must I slow down more? Will the pedestrian on the sidewalk (in orange) continue on the sidewalk or will it also cross the street?

This important task comes with many challenges. First of all, the future is not fully predictable. There are typically multiple possible outcomes, some of them being more likely than others. The controller of a car must be aware of these multiple possibilities and their likelihoods. If a car crashes into a pedestrian who predictably crosses the street, this will be considered a severe failure, whereas extremely unlikely behaviour, such as the pedestrian in purple turning around and crossing the street in the opposite direction, must be ignored to enable efficient control. The approach

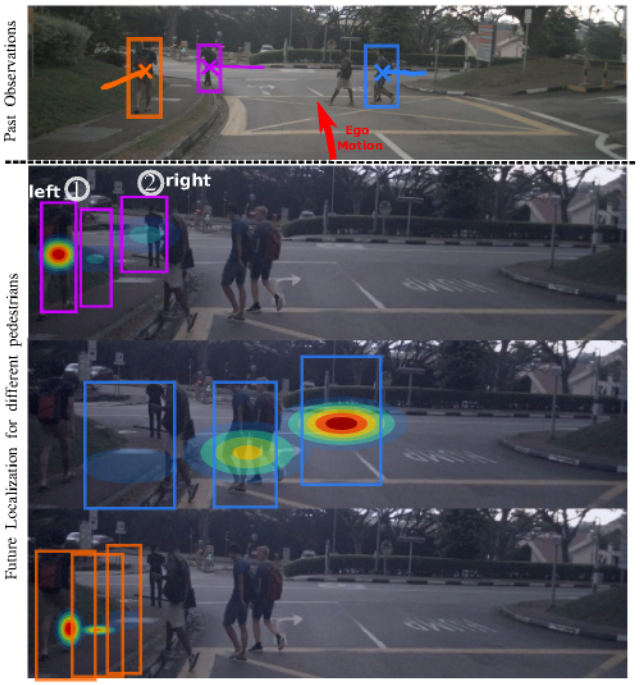


Figure 1. An example from the nuScenes dataset [9]. Given the past observations of pedestrians (colored bounding boxes (top)) and the egomotion of the car (red arrow), our framework predicts multiple modes of their future visualized by a set of bounding boxes and their distribution as an overlaid heatmap. Prediction covers possible options for (2nd row) turning left or right, (3rd row) slowing down or accelerating, (4th row) being on the sidewalk.

we propose predicts two likely modes for this pedestrian: continuing left or right on the sidewalk.

Ideally this task can be accomplished directly in the sensor data without the requirement of privileged information such as a third person view, or a street map that marks all lanes, sidewalks, crossings, etc.. Independence of such information helps the approach generalize to situations not covered by maps or extra sensors, e.g., due to changes not yet captured in the map or GPS failures. However, mak-

ing predictions in egocentric views suffers from partial visibility: we only see the context of the environment in the present view - other relevant parts of the environment are occluded and only become visible as the car moves. Figure 1 shows that the effect of the egomotion is substantial even in this example with relatively slow motion.

In this paper, we approach these two challenges in combination: multimodality of the future and egocentric vision. For the multimodality, we build upon the recent work by Makansi et al. [36], who proposed a technique to overcome mode collapse and stability issues of mixture density networks. However, the work of Makansi et al. assumes a static bird's-eye view of the scene. In order to carry the technical concept over to the egocentric view, we introduce an intermediate prediction which improves the quality of the multimodal distribution: a *reachability prior*. The reachability prior is learned from a large set of egocentric views and tells where objects of a certain class are likely to be in the image based on the image's semantic segmentation; see Figure 2 top. This prior focuses the attention of the prediction based on the environment. Even more important, we can propagate this prior much more easily into the future - using the egomotion of the vehicle - than a whole image or a semantic map. The reachability prior is a condensation of the environment, which contains the semantic context most relevant to the task.

The proposed framework of estimating and propagating a multimodal reachability prior is not only beneficial for future localization of a particular object (Figure 2 left), but it also enables the task of emergence prediction (Figure 2 right). For safe operation, it is not sufficient to reason about the future location of the *observed* objects, but also potentially emerging objects in the scene must be anticipated, if their emergence exceeds a certain probability. For example, passing by a school requires extra care since the probability that a child can jump on the street is higher. Autonomous systems should behave differently near a school exit than on a highway. Predicting emergence of new objects did not yet draw much attention in literature.

The three tasks in Fig. 2 differ via their input conditions: the reachability prior is only conditioned by the semantic segmentation of the environment and the class of interest. It is independent of a particular object. Future localization includes the additional focus on an object of interest and its past trajectory. These conditions narrow down the space of solutions and make the output distribution much more peaked. Emergence prediction is a reduced case of the reachability prior, where new objects can only emerge from unobserved areas of the scene.

In this paper (1) we propose a future localization framework in egocentric view by transferring the work by Makansi et al. [36] from bird's-eye view to egocentric observations, where multimodality is even more difficult to capture. Thus, (2) we propose to compute a reachability prior as intermediate result, which serves as attention to prevent forgetting rare modes, and which can be used to efficiently propagate scene priors into the future taking into account the egomotion. For the first time, (3) we formulate the problem of object emergence prediction for egocentric view with multimodality. (4) We evaluate our approach and the existing methods on the recently largest public nuScenes dataset [9] where the proposed approach shows clear improvements over the state of the art. In contrast to most previous works, the proposed approach is not restricted to a single object category. (5) We include heterogeneous classes like pedestrians, cars, buses, trucks and tricycles. (6) The prediction horizon was tripled from 1 second to 3 seconds into the future compared to existing methods. Moreover, (7) we show that the approach allows zero-shot transfer to unseen and noisy datasets (Waymo [50] and FIT).

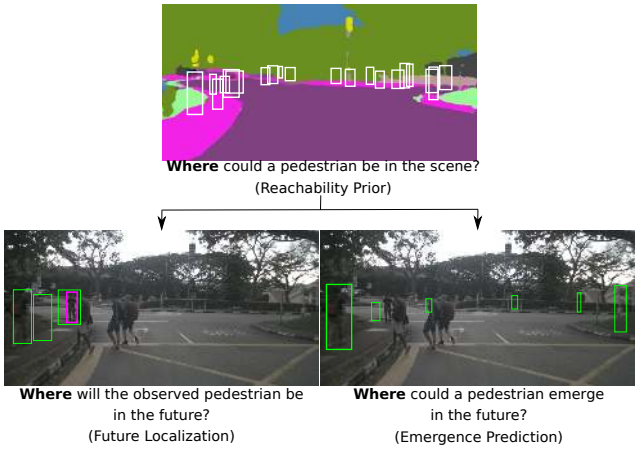


Figure 2. Top: The reachability prior (white rectangles) answers the general question of where a pedestrian could be in a scene. Left: Future localization (green rectangles) of a particular pedestrian crossing the street narrows down the solution from the reachability prior by conditioning the solution on past and current observations. The true future is shown as purple box. Right: The emergence prediction (green rectangles) shows where a pedestrian could suddenly appear and narrows down the solution from the reachability prior by conditioning the solution on the current observation of the scene.

2. Related Work

Bird's-Eye View Future Localization. Predicting the future locations or trajectories of objects is a well studied problem. It includes techniques like the Kalman filter [26], linear regression [39], and Gaussian processes [42, 57, 43, 56]. These techniques are limited to low-dimensional data, which excludes taking into account the semantic context provided by an image. Convolutional networks allow processing such inputs and using them for future localization.

LSTMs have been very popular due to time series processing. Initial works exploited LSTMs for trajectories to model the interaction between objects [2, 59, 65], for scenes to exploit the semantics [4, 38], and LSTMs with attention to focus on the relevant semantics [46].

Another line of works tackle the multimodal nature of the future by sampling through cVAEs [30], GANs [3, 21, 45, 66, 27], and latent decision distributions [31]. Choi et al. [12] model future locations as nonparametric distribution, which can potentially result in multimodality but often collapses to a single mode. Given the instabilities of Mixture Density Networks (MDNs) in unrestricted environments, some works restrict the solution space to a set of predefined maneuvers or semantic areas [15, 24]. Makansi et al. [36] proposed a method to learn mixture densities in unrestricted environments. Their approach first predicts diverse samples and then fits a mixture model on these samples. All these methods have been applied on static scenes recorded from a bird's-eye view, i.e., with full local observability and no egomotion. We build on the technique from Makansi et al. [36] to estimate multimodal distributions in egocentric views.

Egocentric Future Localization. The egocentric camera view is the typical way of observing the scene in autonomous driving. It introduces new challenges due to the egomotion and the narrow field of view. Multiple works have addressed these challenges by projecting the view into bird's-eye view using 3D sensors [14, 17, 16, 47, 35, 44, 13]. This is a viable approach, but it suffers from non-dense measurements or erroneous measurements in case of LIDAR and stereo sensors, respectively.

Alternative approaches try to work directly in the egocentric view. Yagi et al. [62] utilized the pose, locations, scales and past egomotion for predicting the future trajectory of a person. TraPHic [10] exploits the interaction between nearby heterogeneous objects. DTP [49] and STED [48] use encoder-decoder schemes using optical flow and past locations and scales of the objects. Yao et al. [63] added the planned egomotion to further improve the prediction. For autonomous driving, knowing the planned motion is a reasonable assumption [20], and we also make use of this assumption. All these models work with a deterministic model and fail to account for the multimodality and uncertainty of the future. The effect of this is demonstrated by our experiments.

The most related work to our approach, in the sense that it works on egocentric views and predicts multiple modes, is the Bayesian framework by Bhattacharyya et al. [7]. It uses Bayesian RNNs to sample multiple futures with uncertainties. Additionally, they learn the planned egomotion and fuse it to the main future prediction framework. NEMO [37] extends this approach by learning a multimodal distribution for the planned egomotion leading to better accuracy. Both methods need multiple runs to sample different futures and suffer from mode collapse, i.e., tend to predict only the most dominant mode, as demonstrated by our experiments.

Egocentric Emergence Prediction. To reinforce safety in autonomous driving, it is important to not only predict the future of the observed objects but also predict where new objects can emerge. Predicting the whereabouts of an emerging object inherits predicting the future environment itself. Predicting the future environment was addressed by predicting future frames [55, 52, 51, 1, 28, 32, 61] and future semantic segmentation [34, 25, 54, 33, 8, 6]. These methods can only hallucinate new objects in the scene in a photorealistic way, but none of them explicitly predicts the structure where new objects can actually emerge. Vondrick et al. [53] consider a higher-level task and predict the probability of a new object to appear in an egocentric view. However, they only predict "what" object to appear but not "where". Fan et al. [19] suggested transferring current object detection features to the future. This way they anticipate both observed and new objects.

Reachability Prior Prediction. The environment poses constraints for objects during navigation. While some recent works use an LSTM to learn environment constraints from images [38, 60], others [4, 12] choose a more explicit approach by dividing the environment into meaningful grids to learn the grid-grid, object-object and object-grid interactions. Also soft attention mechanisms are commonly used to focus on relevant features of the environments [45, 46]. While these methods reason about static environment constraints within the model proposed, we propose to separate this task and learn a scene prior before the future localization in dynamic scenes. Lee et al [29] proposed a similar module, where a GAN per object class generates multiple locations to place an object photorealistically.

3. Multimodal Egocentric Future Prediction

Figure 3 shows the pipeline of our framework for the future localization task consisting of three main modules: (1) reachability prior network (RPN), which learns a prior of where members of an object class could be located in semantic map, (2) reachability transfer network (RTN), which transfers the reachability prior from the current to a future time step taking into account the planned egomotion, and (3) future localization network (FLN), which is conditioned on the past and current observations of an object and learns to predict a multimodal distribution of its future location based on the general solution from the RTN.

Emergence prediction shares the same first two modules and differs only in the third network where we drop the condition on the past object trajectory. We refer to it as emergence prediction network (EPN). The aim of EPN is to learn a multimodal distribution of where objects of a class emerge in the future.

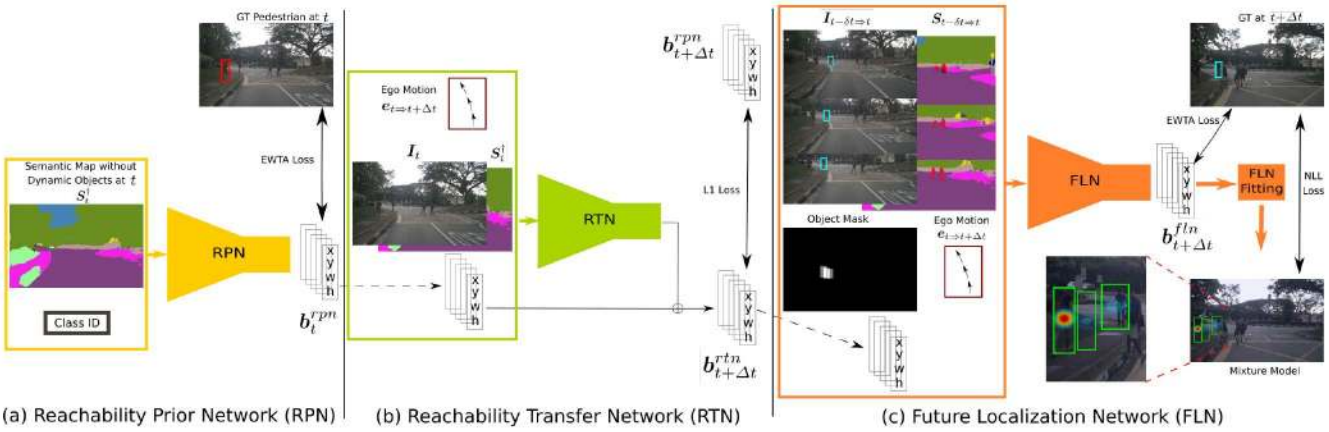


Figure 3. Overview of the overall future localization framework. (a) The reachability prior network (RPN) learns the relation between objects of a certain class ID and the static elements of a semantic map by generating multiple bounding box hypotheses. (b) The reachability transfer network (RTN) transfers the reachability prior into the future given the observed image, its semantic, and the planned egomotion. The ground truth for training this network is obtained in a self-supervised manner by running RPN on the future static semantic map. (c) The future localization network (FLN) yields a multimodal distribution of the future bounding boxes of the object of interest through a sampling network (to generate multiple bounding boxes (samples)) and then a fitting network to fit the samples to a Gaussian mixture model (shown as heatmap overlayed on the future image with the means of the mixture components shown as green bounding boxes). The emergence prediction network (EPN) is identical to the FLN, except that it lacks the object-of-interest masks in the input.

3.1. Reachability Prior Network (RPN)

Given an observed scene from an egocentric view, the reachability prior network predicts where an object of a certain class can be at the same time step in the form of bounding box hypotheses. Let $\boldsymbol{b}_{i,t}^{rpn} = [x, y, w, h]$ for $i \in [1, N]$ be the set of bounding box hypotheses predicted by our RPN at time step t, where (x, y) represents the center coordinates and (w, h) the width and height.

Since the reachability prior network should learn the relation between a class of objects (e.g, vehicle) and the scene semantics (e.g, road, sidewalk, and so on), we remove all dynamic objects from the training samples. This is achieved by inpainting [64]. Because inpainting on the semantic map causes fewer artifacts, in contrast to inpainting in the raw RGB image [5], the reachability prior is based on the semantic map. On one hand, the semantic map does not show some of the useful details visible in the raw image (e.g. the type of traffic sign or building textures). On the other hand, it is important that the inpainting does not introduce strong artifacts. These would be picked up during training and would bias the result (similar to keeping the original objects in the image).

For each image $\boldsymbol{I}_t$ at time t, we compute its semantic segmentation $\boldsymbol{S}_t$ using deeplabV3plus [11] and derive its static semantic segmentation $\boldsymbol{S}_t^{\dagger}$ after inpainting all dynamic objects. This yields the training data for the reachability prior network: the static semantic segmentation is the input to the network, and the removed objects of class c are ground-truth samples for the reachability. The network yields multiple hypotheses $\boldsymbol{b}_{i,t}^{rpn}$ as output and is trained using the EWTA scheme [36] with the loss:

$$L_{RPN} = l(\boldsymbol{b}_{i,t}^{rpn}, \hat{\boldsymbol{b}}_t). \quad (1)$$

$\hat{\boldsymbol{b}}_t$ denotes a ground-truth bounding box of one instance from class c (e.g, vehicle or pedestrian) in image $\boldsymbol{I}_t$ and $l(.)$ denotes the L_2 norm. EWTA applies this loss to the hypotheses in a hierarchical way. It penalizes all hypotheses (i.e, $i \in [1, N]$ where $N = 20$). After convergence, it halves the hypotheses ($N = 10$) and penalizes only the best 10 hypotheses. This halving is repeated until only the best hypothesis is penalized; see Makansi et al. [36] for details. A sample output of the reachability prior network for a car is shown in Figure 4 (top).

3.2. Reachability Transfer Network (RTN)

When running RPN on the semantic segmentation at time t, we obtain a solution for the same time step t. However, at test time, we require this prior in the unobserved future. Thus, we train a network to transfer the reachability at time t to time $t + \Delta t$, where Δt is the fixed prediction horizon and $\boldsymbol{e}_{t\Rightarrow t+\Delta t}$ is the relative pairwise transformation between the pose at time t and $t + \Delta t$ (referred to as planned egomotion) which is represented as a transformation vector (3d translation vector $[t_x, t_y, t_z]$ and rotation quaternions $[q_w, q_x, q_y, q_z]$). This transfer network can be learned with a self-supervised loss from a time series

$$L_{RTN} = \sum_{i=1}^{N} |\boldsymbol{b}_{i,t+\Delta t}^{rtn} - \boldsymbol{b}_{i,t+\Delta t}^{rpn}|. \quad (2)$$

Figure 4. An example from the Waymo [50] dataset showing the reachability prior for the class car in the current time step (RPN: top), the reachability prior transferred to the future (RTN: middle), and final future localization further conditioned on a specific instance (FLN: bottom). For clarity, we draw the hypotheses on both image and semantic domains. Note that **none** of our networks has access to the future image or its semantic map (at time $t + \Delta t$).

where $\boldsymbol{b}^{rtn}_{i,t+\Delta t} = RTN(\boldsymbol{b}^{rpn}_{i,t}, \boldsymbol{e}_{t \Rightarrow t+\Delta t}, \boldsymbol{I}_t, \boldsymbol{S}^{\dagger}_t)$ is the output of the RTN network. $\boldsymbol{I}_t$ is the image and $\boldsymbol{S}^{\dagger}_t$ is the static semantic segmentation at time t. Figure 4 (middle) shows the reachability prior (top) transferred to the future. Given the ego motion as moving forward (red arrow) and the visual cues for upcoming traffic light and a right turn, the RTN anticipates that some more cars can be on the street emerging and transforms some of the RPN hypotheses to cover these new locations.

3.3. Future Localization Network (FLN)

Given an object which is observed for a set of frames from $t - \delta t$ to t, where δt denotes the observation period, FLN predicts the distribution of bounding boxes in the future frame $t + \Delta t$. Figure 3c shows the input to this network: the past images $(\boldsymbol{I}_{t-\delta t}, ..., \boldsymbol{I}_t)$, the past semantic maps $(\boldsymbol{S}_{t-\delta t}, ..., \boldsymbol{S}_t)$, the past masks of the object of interest $(\boldsymbol{M}_{t-\delta t}, ..., \boldsymbol{M}_t)$, the planned egomotion $(\boldsymbol{e}_{t \Rightarrow t+\Delta t})$, and the reachability prior in the future frame $(\boldsymbol{b}^{rtn}_{i,t+\Delta t})$. The object masks $\boldsymbol{M}$s are provided as images, where pixels inside the object bounding box are object class c and 0 elsewhere.

We use the sampling-fitting framework from Makansi et al. [36] to predict a Gaussian mixture for the future bounding box of the object of interest. The sampling network generates multiple hypotheses and is trained with EWTA, just like the RPN. The additional fitting network estimates the parameters (π_k, μ_k, σ_k) of a Gaussian mixture model with $K = 4$ from these hypotheses, similar to the expectation-maximization algorithm but via a network; see Makansi et al. [36] for details. An example of the FLN prediction is shown in figure 4 (bottom). The fitting network is trained with the negative log-likelihood (NLL) loss

$$\boldsymbol{L}_{nll} = -\log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}^2_k) \right] . \quad (3)$$

3.4. Emergence Prediction Network (EPN)

Rather than predicting the future of a seen object, the emergence prediction network predicts where an unseen object can emerge in the scene. The EPN is very similar to the FLN shown in figure 3c. The only difference is that the object masks are missing in the input, since the task is not conditioned on a particular object but predicts the general distribution of objects emerging.

The network is trained on scenes where an object is visible in a later image $\boldsymbol{I}_{t+\Delta t}$ (ground truth), but not in the current image $\boldsymbol{I}_t$. Like for the future localization network, we train the sampling network with EWTA and the fitting network with NLL.

4. Experiments

4.1. Datasets

Mapillary Vistas [41]. We used the Mapillary Vistas dataset for training the inpainting method from [64] on semantic segmentation and for training our reachability prior network. This dataset contains around $25K$ images recorded in different cities across 6 continents, from different viewpoints, and in different weather conditions. For each image, pixelwise semantic and instance segmentation are provided. The images of this dataset are not temporally ordered, which prevents its usage for training the RTN, FLN, or EPN.

nuScenes [9]. nuScenes is very large autonomous driving dataset consisting of 1000 scenes with 20 seconds each. We used it for training and evaluating the proposed framework. We did *not* re-train the reachability prior network on this dataset, as to test generalization of the reachability prior network across different datasets. The nuScenes dataset provides accurate bounding box tracking for different types of traffic objects and the egomotion of the observer vehicle. We used the standard training/validation split (700/150 scenes) of the dataset for training/evaluating all experiments.

Waymo Open Dataset [50]. Waymo is the most recent autonomous driving dataset and contains 1000 scenes with 20 seconds each. To show zero-shot transfer of our framework (i.e, without re-training the model), we used the standard 202 testing scenes.

FIT Dataset. We collected 18 scenes from different locations in Europe and relied on MaskRCNN [22] and deep-

sort [58] to detect and track objects, and DSO [18] to estimate the egomotion. This dataset allows testing the robustness to noisy inputs (without human annotation). We will make these sequences and the annotations publicly available.

4.2. Evaluation Metrics

FDE. For evaluating both future localization and emergence prediction, we report the common Final Displacement Error (FDE), which estimates the L_2 distance of the centers of two bounding boxes in pixels.

IOU. We report the Intersection Over Union (IOU) metric to evaluate how well two bounding boxes overlap.

The above metrics are designed for single outputs, not distributions. In case of multiple hypotheses, we applied the above metrics between the ground truth and the closest mode to the ground truth (known as *Oracle* [36, 30]).

NLL. To evaluate the accuracy of the multimodal distribution, we compute the negative log-likelihood of the ground-truth samples according to the estimated distribution.

4.3. Training Details

We used ResNet-50 [23] as sampling network in all parts of this work. The fitting network consisted of two fully connected layers (each with 500 nodes) with a dropout layer (rate = 0.2) in between. In the FLN, we observed $\delta t = 1$ second and predicted $\Delta t = 3$ seconds into the future. For the EPN, we observed only one frame and predicted $\Delta t = 1$ second into the future. We used $N = 20$ for all sampling networks, and $K = 4$ and $K = 8$ as the number of mixture components for the FLN and the EPN, respectively. The emergence prediction task requires more modes compared to the future localization task since the distribution has typically more modes in this task.

4.4. Baselines

As there is only one other work so far on egocentric multimodal future prediction [7], we compare also to unimodal baselines, which are already more established.

Kalman Filter [26]. This linear filter is commonly used for estimating the future state of a dynamic process through a set of (low-dimensional) observations. It is not expected to be competitive, since it considers only the past trajectory and ignores all other information.

DTP [49]. DTP is a dynamic trajectory predictor for pedestrians based on motion features obtained from optical flow. We used their best performing framework, which predicts the difference to the constant velocity solution.

STED [48]. STED is a spatial-temporal encoder-decoder that models visual features by optical flow and temporal features by the past bounding boxes through GRU encoders. It later fuses the encoders into another GRU decoder to obtain the future bounding boxes.

RNN-ED-XOE [63]. RNN-ED-XOE is an RNN-based encoder-decoder framework which models both temporal and visual features similar to STED. RNN-ED-XOE additionally encodes the future egomotion before fusing all information into a GRU decoder for future bounding boxes.

FLN-Bayesian using [7]. The work by Bhattacharyya et al. [7] is the only multimodal future prediction work for the egocentric scenario in the literature. It uses Bayesian optimization to estimate multiple future hypotheses and their uncertainty. Since they use a different network architecture and data modalities, rather than direct method comparison we port their Bayesian optimization into our framework for fair comparison. We re-trained our FLN with their objective to create samples by dropout during training and testing time as replacement for the EWTA hypotheses. We used the same number of samples, $N = 20$, as in our standard approach.

All these baselines predict the future trajectory of either pedestrians [49, 48, 7] or vehicles [63]. Thus, we re-trained them on nuScenes [9] to handle both pedestrian and vehicle classes. Moreover, some baselines utilize the future egomotion obtained from ORB-SLAM2 [40] or predicted by their framework, as in [7]. For a fair comparison, we used the egomotion from nuScenes dataset when re-training and testing their models, thus eliminating the effect of different egomotion estimation methods.

FLN w/o reachability. To measure the effect of the reachability prior, we ran this version of our framework without RPN and RTN.

FLN + reachability. Our full framework including all 3 networks: RPN, RTN, FLN.

Due to the lack of comparable work addressing the emergence prediction task, so far, we conduct an ablation study on the emergence prediction to analyze the effect of the proposed reachability prior on the accuracy of the prediction.

4.5. Egocentric Future Localization

Table 1 shows a quantitative evaluation of our proposed framework against all the baselines listed above. To distinguish test cases that can be solved with simple extrapolation from more difficult cases, we use the performance of the Kalman filter [26]; see also [63]. A test sample, where the Kalman filter [26] has a displacement error larger than average is counted as *challenging*. An error more than twice the average is marked *very challenging*. In Table 1, we show the error only for the whole test set (all) and the very challenging subset (hard). More detailed results are in the supplemental material.

As expected, deep learning methods outperform the extrapolation by a Kalman filter on all metrics. Both variants of our framework show a significant improvement over all

	nuScenes [9] (all $11k$ / hard $1.4k$)			Waymo [50] (all $47.2k$ / hard $7.1k$)			FIT (all $1.4k$ / hard 223)		
	FDE ↓	IOU ↑	NLL ↓	FDE ↓	IOU ↑	NLL ↓	FDE ↓	IOU ↑	NLL ↓
Kalman [26]	45.02/179.92	0.31/0.01	–	31.69/124.71	0.39/0.02	–	38.33/146.50	0.36/0.03	–
DTP [49]	35.88/111.49	0.34/0.05	–	28.31/ 82.64	0.38/0.10	–	34.99/118.36	0.37/0.09	–
RNN-ED-XOE [63]	30.47/ 78.54	0.34/0.13	–	25.23/ 59.23	0.36/0.18	–	35.74/ 88.58	0.36/0.17	–
STED [48]	27.71/ 82.71	0.39/0.13	–	20.73/ 58.14	0.42/0.20	–	31.80/ 86.58	0.35/0.16	–
FLN-Bayesian using [7]	28.51/ 82.23	0.37/0.13	19.75/28.44	23.75/ 64.67	0.38/0.17	18.80/27.54	32.64/ 87.63	0.38/0.16	20.56/28.83
FLN w/o RPN	15.91/ 47.15	0.54/0.29	19.46/26.85	13.20/ 36.57	0.54/0.34	18.84/26.19	18.12/ 47.92	0.53/0.33	20.38/27.88
FLN + RPN	**12.82**/ **32.68**	**0.55**/**0.33**	**17.90**/**24.17**	**10.35**/ **27.15**	**0.58**/**0.37**	**16.63**/**22.95**	**15.41**/ **32.14**	**0.54**/**0.39**	**19.08**/**24.73**

Table 1. Result for future localization on the nuScenes [9], the Waymo [50], and our FIT datasets. The bottom three methods predict a multimodal distribution. The other methods are not probabilistic and do not allow evaluation of the NLL. For each cell, we report the average over (all testing scenarios/the very challenging scenarios). The number of all/very challenging scenarios for each dataset is shown in parentheses (top).

baselines for the FDE and IOU metrics. When we use FDE or IOU, we use the oracle selection of the hypotheses (i.e, the closest bounding box to the ground truth). Hence, a multimodal method is favored over a unimodal one. Still, such significant improvement indicates the need for multimodality. To evaluate without the bias introduced by the oracle selection, we also report the negative log-likelihood (NLL).

Both variants of the proposed framework outperform the Bayesian framework on all metrics including the NLL. In fact, the Bayesian baseline is very close to the best unimodal baseline. This indicates its tendency for mode collapse, which we also see qualitatively. The use of the reachability prior is advantageous on all metrics and for all difficulties.

As the networks (ours and all baselines) were trained on nuScenes, the results on Waymo and FIT include a zero-shot transfer to unseen datasets. We obtain the same ranking for unseen datasets as for the test set of nuScenes. This indicates that overfitting to a dataset is not an issue for this task. We recommend having cross-dataset experiments (as we show) also in future works to ensure that this stays true and future improvements in numbers are really due to better models and not just overfitting.

Figure 5 shows some qualitative example in four challenging scenarios, where there are multiple options for the future location. (1) A pedestrian starts crossing the street and his future is not deterministic due to different speed estimates. (2) A pedestrian enters the scene from the left and will either continue walking to cross the street or will stop at the traffic light. (3) A tricycle driving from a parking area will continue driving to cross the road or will stop to give way to our vehicle. (4) A car entering the scene from the left will either slow down to yield or drive faster to overpass.

For all scenarios, we observe that the reachability prior (shown as set of colored bounding boxes) defines the general relation between the object of interest and the static elements of the scene. Similar to the observation from our quantitative evaluation, the Bayesian baseline predicts a single future with some uncertainty (unimodal distribution). Our framework without exploiting the reachability prior (FLN w/o RPN) tends to predict more diverse futures

	FDE ↓	IOU ↑	NLL ↓
EPN w/o RPN	21.48	0.18	22.99
EPN + RPN	**15.89**	**0.19**	**21.03**

Table 2. Quantitative results for the emergence prediction task on the nuScenes dataset [9].

but still lacks predicting many of the modes. The reachability prior helps the approach to cover more of the possible future locations.

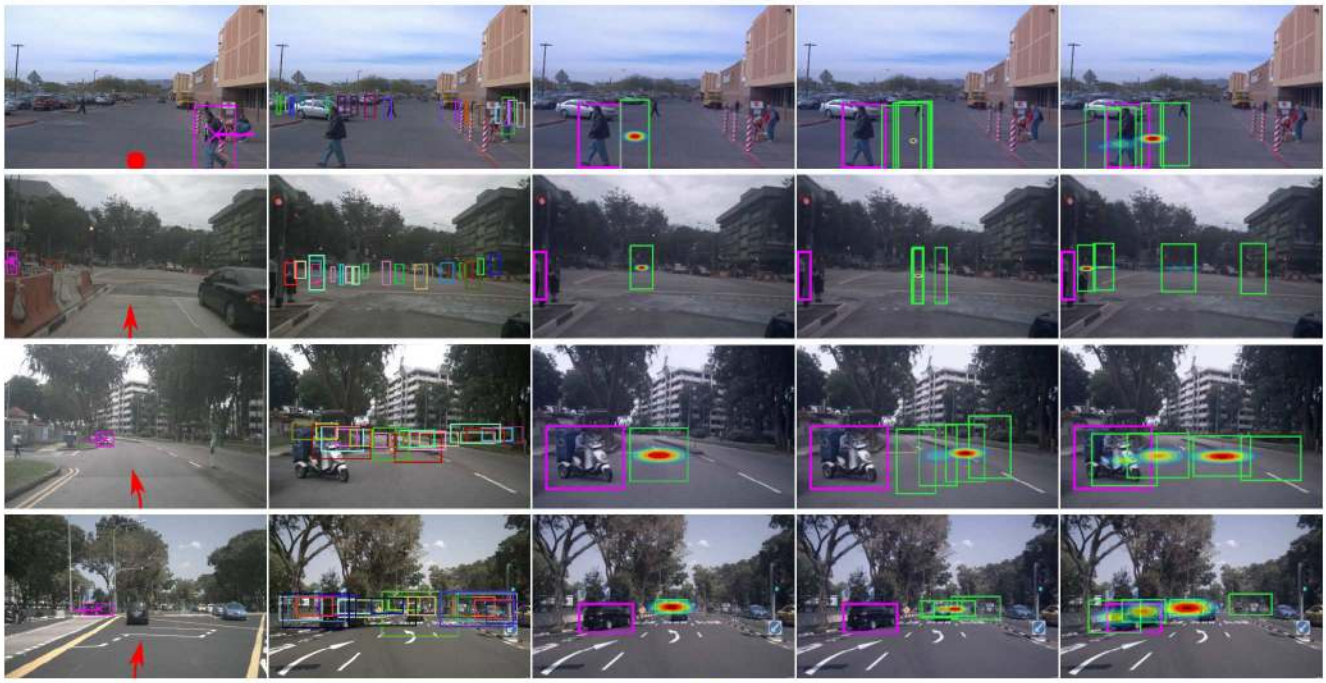
We highly recommend watching the supplementary video, which gives a much more detailed qualitative impression of the results, as it allows the observer to get a much better feeling for the situation than the static pictures in the paper.

4.6. Egocentric Emergence Prediction

Table 2 shows the ablation study on the importance of using the reachability prior for the task of predicting object emergence in a scene. Similar to future localization, exploiting the reachability prior yields a higher accuracy and captures more of the modes. Two qualitative examples for this task are shown in Figure 6. Examples include scenarios (1) where a vehicle could emerge in the scene from the left street, could pass by or could be oncoming; (2) where a car could emerge from the left, from the right, it could pass by, or could be oncoming. EPN learns not only the location in the image, but also meaningful scales. For instance, the anticipation of passing-by cars has a larger scale compared to expected oncoming cars. The distributions for the two examples are different since more modes for emerging vehicles are expected in the second example (e.g, emerging from the right side). Notably, the reachability prior solution is different from the emergence solution, where close-by cars in front of the egocar are part of the reachability prior solution but are ruled out, since a car cannot suddenly appear there. More results are provided in the supplemental material.

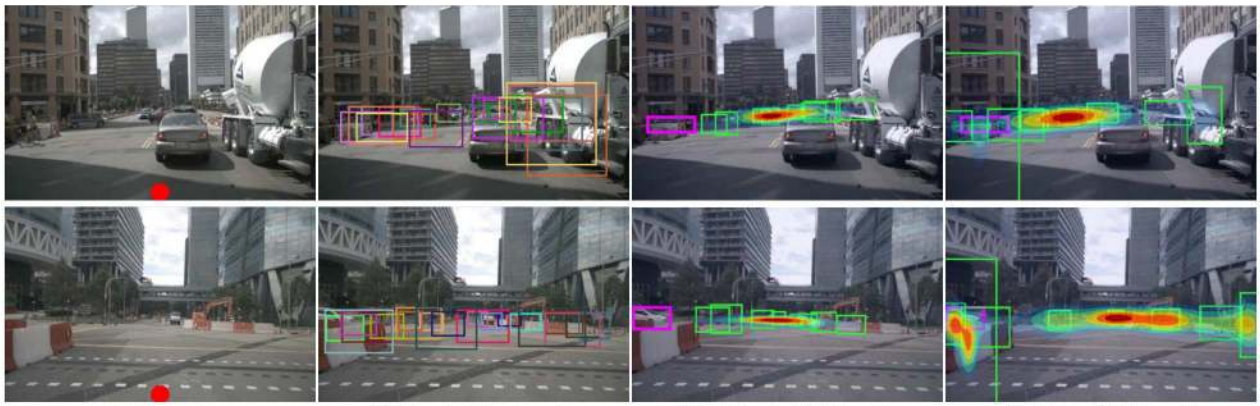
5. Conclusions

In this work, we introduced a method for predicting future locations of traffic objects in egocentric views without



(a) Input $t - \delta t \Rightarrow t$ (b) Reachability $t + \Delta t$ (c) FLN-Bayesian [7] $t + \Delta t$ (d) FLN w/o RPN $t + \Delta t$ (e) FLN w RPN $t + \Delta t$

Figure 5. Results for future localization on Waymo [50] (1st row) and nuScenes [9] (2-4 rows). For each row (scenario), we show (a) the observed trajectory of the object of interest (pink) and the planned egomotion (red arrow) to the future (red circle indicates no egomotion), (b) the reachability prior resulted from the RTN in the future frame, (c) a heatmap overlayed on the future image and the mean prediction (green bounding box) visualizing the distribution predicted by the Bayesian method and the ground-truth bounding box (pink), (d-e) both variants of our future localization framework.



(a) Input t (b) Reachability $t + \Delta t$ (d) EPN w/o RPN $t + \Delta t$ (e) EPN w RPN $t + \Delta t$

Figure 6. Sample results for emergence prediction on nuScenes [9]. For each row (scenario), we show (a) the observed image and the planned egomotion to the future (red circle indicates no egomotion), (b) the reachability prior from the RTN in the future frame, (c-d) both variants of the emergence prediction framework.

predefined assumptions on the scene and by taking into account the multimodality of the future. We showed that a reachability prior and multi-hypotheses learning help overcome mode collapse. We also introduced a new task relevant for autonomous driving: predicting locations of suddenly emerging objects. Overall, we obtained quite good results even in difficult scenarios, but careful qualitative inspection of many results still shows a lot of potential for improvement on future prediction.

6. Acknowledgments

This work was funded in parts by IMRA Europe S.A.S. and the German Ministry for Research and Education (BMBF) via the project Deep-PTL.

References

[1] Sandra Aigner and Marco Krner. Futuregan: Anticipating the future frames of video sequences using spatio-temporal 3d convolutions in progressively growing gans. *arXiv preprint arXiv:1810.01325*, 2018.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.

[3] Javad Amirian, Jean-Bernard Hayet, and Julien Pettre. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *CVPR Workshops*, 2019.

[4] Federico Bartoli, Giuseppe Lisanti, Lamberto Ballan, and Alberto Del Bimbo. Context-aware trajectory prediction. *arXiv preprint arXiv:1705.02503*, 2017.

[5] Lorenzo Berlincioni, Federico Becattini, Leonardo Galteri, Lorenzo Seidenari, and Alberto Del Bimbo. Semantic road layout understanding by generative adversarial inpainting. *arXiv preprint arXiv:1805.11746*, 2018.

[6] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Bayesian prediction of future street scenes through importance sampling based optimization. *arXiv preprint arXiv:1806.06939*, 2018.

[7] A. Bhattacharyya, M. Fritz, and B. Schiele. Long-term on-board prediction of people in traffic scenes under uncertainty. In *CVPR*, June 2018.

[8] Apratim Bhattacharyya, Mario Fritz, and Bernt Schiele. Bayesian prediction of future street scenes using synthetic likelihoods. In *ICLR*, 2019.

[9] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.

[10] Rohan Chandra, Uttaran Bhattacharya, Aniket Bera, and Dinesh Manocha. Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions. In *CVPR*, June 2019.

[11] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[12] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *ICCV*, 2019.

[13] Chiho Choi, Abhishek Patil, and Srikanth Malla. Drogon: A causal reasoning framework for future trajectory forecast. *arXiv preprint arXiv:1908.00024*, 2019.

[14] Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Jeff Schneider, David Bradley, and Nemanja Djuric. Deep kinematic models for physically realistic prediction of vehicle trajectories. *arXiv preprint arXiv:1908.00219*, 2019.

[15] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. *arXiv preprint arXiv:1809.10732*, 2018.

[16] N. Deo, A. Rangesh, and M. M. Trivedi. How would surround vehicles move? a unified framework for maneuver classification and motion prediction. *T-IV*, June 2018.

[17] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, and Jeff Schneider. Short-term motion prediction of traffic actors for autonomous driving using deep convolutional networks. *arXiv preprint arXiv:1808.05819*, 2018.

[18] J. Engel, V. Koltun, and D. Cremers. Direct sparse odometry. In *arXiv:1607.02565*, 2016.

[19] Chenyou Fan, Jangwon Lee, and Michael S. Ryoo. Forecasting hands and objects in future frames. In *ECCV*, September 2018.

[20] D. Gonzlez, J. Prez, V. Milans, and F. Nashashibi. A review of motion planning techniques for automated vehicles. *T-ITS*, April 2016.

[21] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, 2018.

[22] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, Oct 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016.

[24] Yeping Hu, Wei Zhan, and Masayoshi Tomizuka. Probabilistic prediction of vehicle semantic intention and motion. *arXiv preprint arXiv:1804.03629*, 2018.

[25] Xiaojie Jin, Huaxin Xiao, Xiaohui Shen, Jimei Yang, Zhe Lin, Yunpeng Chen, Zequn Jie, Jiashi Feng, and Shuicheng Yan. Predicting scene parsing and motion dynamics in the future. In *NIPS*, pages 6915–6924, 2017.

[26] R. E. Kalman. A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*, 1960.

[27] Vineet Kosaraju, Amir Sadeghian, Roberto Martn-Martn, Ian Reid, S. Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *arXiv preprint arXiv:1907.03395*, 2019.

[28] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *CVPR*, June 2019.

[29] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. In *NIPS*, 2018.

[30] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *CVPR*, pages 336–345, 2017.

[31] Yuke Li. Which way are you going? imitative decision learning for path forecasting in dynamic scenes. In *CVPR*, June 2019.

[32] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *ECCV*, 2018.

[33] Pauline Luc, Camille Couprie, Yann Lecun, and Jakob Verbeek. Predicting future instance segmentations by forecasting convolutional features. *arXiv preprint arXiv:1803.11496*, 2018.

[34] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun. Predicting deeper into the future of semantic segmentation. In *ICCV*, 2017.

[35] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*, pages 6120–6127, 2019.

[36] Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *CVPR*, June 2019.

[37] Srikanth Malla and Chiho Choi. Nemo: Future object localization using noisy ego priors. *arXiv preprint arXiv:1909.08150*, 2019.

[38] Huynh Manh and Gita Alaghband. Scene-lstm: A model for human trajectory prediction. *arXiv preprint arXiv:1808.04018*, 2018.

[39] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989.

[40] R. Mur-Artal and J. D. Tards. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 2017.

[41] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017.

[42] A. O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):1–24, 1978.

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

[44] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *ICCV*, 2019.

[45] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *CVPR*, 2019.

[46] Amir Sadeghian, Ferdinand Legros, Maxime Voisin, Ricky Vesel, Alexandre Alahi, and Silvio Savarese. Car-net: Clairvoyant attentive recurrent network. In *ECCV*, 2018.

[47] Shashank Srikanth, Junaid Ahmed Ansari, Karnik Ram R, Sarthak Sharma, Krishna Murthy J., and Madhava Krishna K. Infer: Intermediate representations for future prediction. *arXiv preprint arXiv:1903.10641*, 2019.

[48] Olly Styles, Tanaya Guha, and Victor Sanchez. Multiple object forecasting: Predicting future object locations in diverse environments. *arXiv preprint arXiv:1909.11944*, 2019.

[49] O. Styles, A. Ross, and V. Sanchez. Forecasting pedestrian trajectory with machine-annotated training data. In *IV*, June 2019.

[50] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset, 2019.

[51] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017.

[52] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017.

[53] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *CVPR*, pages 98–106, 2016.

[54] Suhani Vora, Reza Mahjourian, Soeren Pirk, and Anelia Angelova. Future segmentation using 3d structure. *arXiv preprint arXiv:1811.11358*, 2018.

[55] J. Walker, A. Gupta, and M. Hebert. Patch to the future: Unsupervised visual prediction. In *CVPR*, 2014.

[56] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *TPAMI*, 30(2):283–298, 2008.

[57] C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning and Inference in Graphical Models*, pages 599–621. Kluwer, 1997.

[58] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. *arXiv preprint arXiv:1703.07402*, 2017.

[59] Y. Xu, Z. Piao, and S. Gao. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In *CVPR*, pages 5275–5284, 2018.

[60] H. Xue, D. Q. Huynh, and M. Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *WACV*, 2018.

[61] Tianfan Xue, Jiajun Wu, Katherine Bouman, and Bill Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *NIPS*, pages 91–99, 2016.

[62] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato. Future person localization in first-person videos. In *CVPR*, June 2018.

[63] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush. Egocentric vision-based future vehicle localization for intelligent driving assistance systems. In *ICRA*, May 2019.

[64] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *CVPR*, June 2018.

[65] Pu Zhang, Wanli Ouyang, Pengfei Zhang, Jianru Xue, and Nanning Zheng. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *CVPR*, 2019.

[66] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *CVPR*, June 2019.