Article

The International Journal of Robotics Research
1–41

DOI: 10.1177/0278364920917446
journals.sagepub.com/home/ijr
SAGE

Human motion trajectory prediction: a survey

Andrey Rudenko[1,2], Luigi Palmieri[1], Michael Herman[3], Kris M Kitani[4], Dariu M Gavrila[5] and Kai O Arras[1]

Abstract

With growing numbers of intelligent autonomous systems in human environments, the ability of such systems to perceive, understand, and anticipate human behavior becomes increasingly important. Specifically, predicting future positions of dynamic agents and planning considering such predictions are key tasks for self-driving vehicles, service robots, and advanced surveillance systems. This article provides a survey of human motion trajectory prediction. We review, analyze, and structure a large selection of work from different communities and propose a taxonomy that categorizes existing methods based on the motion modeling approach and level of contextual information used. We provide an overview of the existing datasets and performance metrics. We discuss limitations of the state of the art and outline directions for further research.

Keywords

Survey, review, motion prediction, robotics, video surveillance, autonomous driving

1. Introduction

Understanding human motion is a key skill for intelligent systems to coexist and interact with humans. It involves aspects in representation, perception, and motion analysis. Prediction plays an important part in human motion analysis: foreseeing how a scene involving multiple agents will unfold over time allows to incorporate this knowledge in a pro-active manner, i.e., allowing for enhanced ways of active perception, predictive planning, model predictive control, or human–robot interaction. As such, human motion prediction has received increased attention in recent years across several communities. Many important application domains exist, such as self-driving vehicles, service robots, and advanced surveillance systems, see Figure 1.

The challenge of making accurate predictions of human motion arises from the complexity of human behavior and the variety of its internal and external stimuli. Motion behavior may be driven by own goal intent, the presence and actions of surrounding agents, social relations between agents, social rules and norms, or the environment with its topology, geometry, affordances, and semantics. Most factors are not directly observable and need to be inferred from noisy perceptual cues or modeled from context information. Furthermore, to be effective in practice, motion prediction should be robust and operate in real-time.

Human motion comes in many forms: articulated full-body motion, gestures and facial expressions, or movement through space by walking, using a mobility device or driving a vehicle. The scope of this survey is human motion trajectory prediction. Specifically, we focus on ground-level 2D trajectory prediction for pedestrians and also consider the literature on cyclists and vehicles. Prediction of video frames, articulated motion, or human actions or activities is out of the scope of this work although many of those tasks rely on the same motion modeling principles and trajectory prediction methods considered here. Within this scope, we survey a large selection of works from different communities and propose a novel taxonomy based on the motion modeling approaches and the contextual cues. We categorize the state of the art and discuss typical properties, advantages and drawbacks of the categories as well as outline open challenges for future research. Finally, we raise three questions.

(*Q1*) Are the evaluation techniques to measure prediction performance good enough and follow best practices?

[1]Robert Bosch GmbH, Corporate Research, Germany
[2]Mobile Robotics and Olfaction Lab, Örebro University, Sweden
[3]Bosch Center for Artificial Intelligence, Germany
[4]Carnegie Mellon University, USA
[5]Intelligent Vehicles group, TU Delft, The Netherlands

Corresponding author:
Andrey Rudenko, Robert Bosch GmbH, Department CR/AAS3, Robert-Bosch-Campus 1, Renningen 71272, Germany.
Email: andrey.rudenko@de.bosch.com

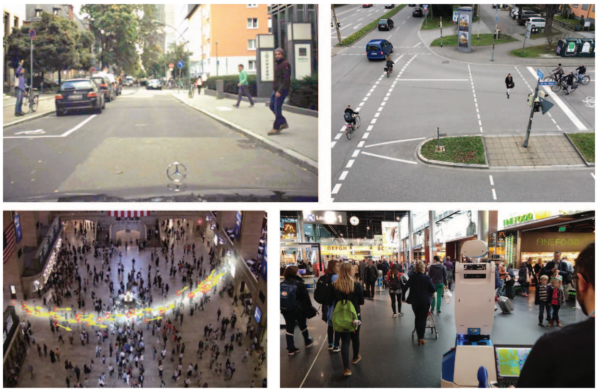

Fig. 1. Application domains of human motion prediction. Top left: Will the pedestrian cross? Self-driving vehicles have to quickly reason about intentions and future locations of other traffic participants, such as pedestrians (illustration from Kooij et al., 2019). Top right: Advanced traffic surveillance systems can provide real-time alerts of pending collisions using communication technology. Bottom left: Advanced surveillance systems analyze human motion in public spaces for suspicious activity detection or crowd control (illustration from Zhou et al., 2015). Bottom right: Robot navigation in densely populated spaces requires accurate motion prediction of surrounding people to safely and efficiently move through crowds.

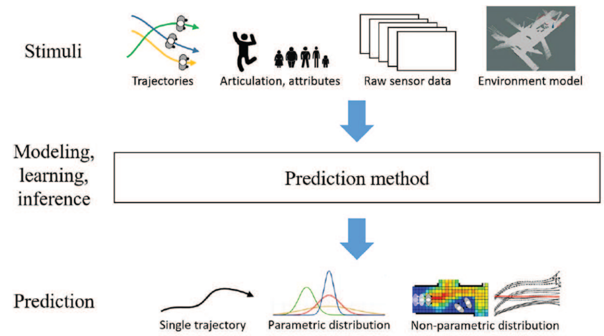

Fig. 2. Typical elements of a motion prediction system: internal and external stimuli that influence motion behavior, the method itself and the different parametric, non-parametric, or structured forms of predictions.

(*Q2*) Have all prediction methods arrived on the same performance level and the choice of the modeling approach does not matter anymore?

(*Q3*) Is motion prediction solved?

The paper is structured as follows: we present the taxonomy in Section 2, review and analyze the literature on human motion prediction first by the modeling approaches in Sections 3–5, and then by the contextual cues in Section 6. In Section 7 we review the benchmarking of motion prediction techniques in terms of commonly used performance metrics and datasets. In Section 8 we discuss the state of the art with respect to the above three questions and outline open research challenges. Finally, Section 9 concludes the paper.

We recommend Sections 1 and 2, Figures 8–10, and Section 8 as a coarse overview of the motion prediction methodology for a general reader. A practitioner may find value in the review of the datasets and metrics in Section 7. Finally, the thorough analysis of the literature in Sections 3–6 is recommended for expert readers.

1.1. Overview and terminology

On the highest level of abstraction, the motion prediction problem contains the following three elements (Figure 2).

- *Stimuli:* Internal and external stimuli that determine motion behavior include the agents' motion intent and other directly or indirectly observable influences. Most prediction methods rely on observed partial trajectories, or generally, sequences of agent state observations such as positions, velocities, body joint angles, or attributes. Often, this is provided by a target tracking system and it is common to assume correct track identity over the observation period. Other forms of inputs include contextual cues from the environment such as scene geometry, semantics, or cues that relate to other moving entities in the surrounding. End-to-end approaches rely on sequences of raw sensor data.
- *Modeling approach:* Approaches to human motion prediction differ in the way they represent, parametrize, learn, and solve the task. This article focuses on finding and analyzing useful categories, hidden similarities, common assumptions, and best evaluation practices in the growing body of literature.
- *Prediction:* Different methods produce different parametric, non-parametric, or structured forms of predictions such as Gaussians over agent states, probability distributions over grids, singular or multiple trajectory samples, or motion patterns using graphical models.

We use the term *agent* to denote dynamic objects of interest such as robots, pedestrians, cyclists, cars, or other human-driven vehicles. The *target agent* is the dynamic object for which we make the actual motion prediction. We assume the agent behavior to be non-erratic and goal-directed with regard to an optimal or near-optimal expected outcome. This assumption is typical as the motion prediction problem were much harder or even ill-posed otherwise. We define a *path* to be a sequence of (x, y)-positions and a *trajectory* to be a path combined with a timing law or a velocity profile. We refer to *short-term* and *long-term* prediction to characterize prediction horizons of 1–2 s and up to 20 s ahead, respectively.

Formally, we use $\mathbf{s}_t$ to denote the state of an agent at time t, $\mathbf{u}_t$ to denote the action that the agent takes at time t, $\mathbf{o}_t \in \mathcal{O}$ to denote the observations of the agent's state at time t, and use ζ to denote trajectories. We refer to a history of several states, actions, or observations from time t to time T using subscripts $t : T$.

1.2. Application domains

Motion prediction is a key task for service robots, self-driving vehicles, and advanced surveillance systems (Figure 1).

1.2.1. Service robots. Mobile service robots increasingly operate in open-ended domestic, industrial, and urban environments shared with humans. Anticipating motion of surrounding agents is an important prerequisite for safe and efficient motion planning and human–robot interaction. Limited on-board resources for computation and first-person sensing makes this a challenging task.

1.2.2. Self-driving vehicles. The ability to anticipate motion of other road users is essential for automated driving. Similar challenges apply as in the service robot domain, although they are more pronounced given the higher masses and velocities of vehicles and the resulting larger harm that can potentially be inflicted, especially towards vulnerable road users (i.e., pedestrians and cyclists). Furthermore, vehicles need to operate in rapidly changing, semantically rich outdoor traffic settings and need hard real-time operating constraints. Knowledge of the traffic infrastructure (location of lanes, curbside, traffic signs, traffic lights, other road markings such as zebra crossings) and the traffic rules can help in motion prediction.

1.2.3. Surveillance. Visual surveillance of vehicular traffic or human crowds relies on the ability to accurately track a large number of targets across distributed networks of stationary cameras. Long-term motion prediction can support a variety of surveillance tasks such as person retrieval, perimeter protection, traffic monitoring, crowd management, or retail analytics by further reducing the number of false-positive tracks and track identifier switches, particularly in dense crowds or across non-overlapping fields of views.

1.3. Related surveys

In this section, we detail related surveys from different scientific communities, i.e., robotics (Chik et al., 2016; Kruse et al., 2013; Lasota et al., 2017), intelligent vehicles (Brouwer et al., 2016; Lefèvre et al., 2014; Ridel et al., 2018), and computer vision (Hirakawa et al., 2018; Morris and Trivedi, 2008; Murino et al., 2017).

Kruse et al. (2013) provided a survey of approaches for wheeled mobile robots and categorized human-aware motion based on comfort, naturalness, and sociability features. Motion prediction is seen as part of a human-aware navigation framework and categorized into *reasoning-based* and *learning-based* approaches. In reasoning-based methods, predictions are based on simple geometric reasoning or dynamic models of the target agent. Learning-based approaches make predictions via motion patterns that are learned from observed agent trajectories.

A short survey on frameworks for socially-aware robot navigation was provided by Chik et al. (2016). The authors discussed key components of such frameworks including several planners and human motion prediction techniques.

Lasota et al. (2017) surveyed the literature on safe human–robot interaction along the four themes of safety through control, motion planning, prediction, and psychological factors. In addition to wheeled robots, they also include related works on manipulator arms, drones, or self-driving vehicles. The literature on human motion prediction is divided into methods based on *goal intent* or *motion characteristics*. Goal intent techniques infer an agent's goal and predict a trajectory that the agent is likely to take to reach that goal. The latter group of approaches does not rely explicitly on goals and makes use of observations about how humans move and plan natural paths.

Lefèvre et al. (2014) surveyed vehicular motion prediction and risk assessment in an automated driving context. The authors discussed the literature based on the semantics used to define motion and risk and distinguish *physics-based*, *maneuver-based*, and *interaction-aware* models for prediction. Physics-based methods predict future trajectories via forward simulation of a vehicle model, typically under kinodynamic constraints and uncertainties in initial states and controls. Maneuver-based methods assume that vehicle motion is a series of typical motion patterns (maneuvers) that have been acquired a priori and can be recognized from observed partial agent trajectories. Intention-aware methods make joint predictions that account for inter-vehicle interactions, also considering that such interactions are regulated by traffic rules.

Brouwer et al. (2016) reviewed and compared pedestrian motion models for vehicle safety systems. According to the cues from the environment used as input for motion prediction, the authors distinguished four classes of methods: *dynamics-based models*, which only use the target agent's motion state; methods that use *psychological knowledge of human behavior* in urban environments (e.g., probabilities of acceleration, deceleration, switch of the dynamical model); methods that use *head orientation*; and methods that use a *semantic map* of the environment. This categorization was extended by Ridel et al. (2018) to review pedestrian crossing intention inference techniques.

Morris and Trivedi (2008) surveyed methods for trajectory learning and analysis for visual surveillance. They discussed similarity metrics, techniques, and models for learning prototypical motion patterns (called activity paths) and briefly considered trajectory prediction as a case of online activity analysis. Murino et al. (2017) discussed group and crowd motion analysis as a multidisciplinary problem that combines insights from the social sciences with concepts

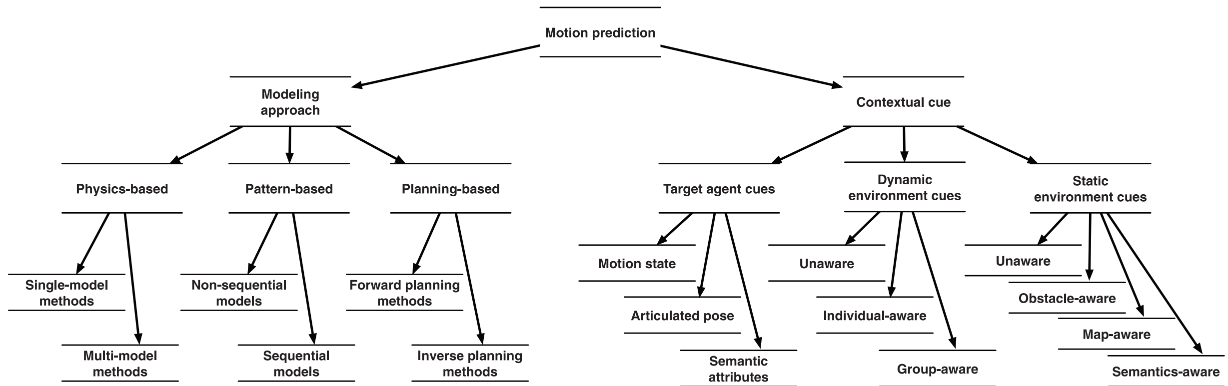

Fig. 3. Overview of the categories in our taxonomy.

from computer vision and pattern recognition. The authors reviewed several recent methods for tracking and prediction of human motion in crowds. Hirakawa et al. (2018) surveyed video-based methods for semantic feature extraction and human trajectory prediction. The literature is divided based on the motion modeling approach into *Bayesian models*, *energy minimization methods*, *deep learning methods*, *inverse reinforcement learning (IRL) methods*, and *other* approaches.

Related to our discussion of the benchmarking practices, several works survey the datasets of motion trajectories (Hirakawa et al., 2018; Poiesi and Cavallaro, 2015; Ridel et al., 2018) and metrics for prediction evaluation (Quehl et al., 2017). Poiesi and Cavallaro (2015) and Hirakawa et al. (2018) described several datasets of human trajectories in crowded scenarios, used to study social interactions and evaluate path prediction algorithms. Ridel et al. (2018) discussed available datasets of pedestrian motion in urban settings. Quehl et al. (2017) reviewed several trajectory similarity metrics, applicable in the motion prediction context.

Unlike these surveys, we review and analyze the literature across multiple application domains and agent types. Our taxonomy offers a novel way to structure the growing body of literature, containing the categories proposed by Kruse et al. (2013), Lasota et al. (2017), and Lefèvre et al. (2014) and extending them with a systematic categorization of contextual cues. In particular, we argue that the modeling approach and the contextual cues are two fundamentally different aspects underlying the motion prediction problem and should be considered separate dimensions for the categorization of methods. This allows, for example, the distinction of physics-based methods that are unaware of any external stimuli from methods in the same category that are highly situational aware accounting for road geometry, semantics, and the presence of other agents. This is unlike previous surveys whose categorizations are along a single dimension based on both different modeling approaches and increasing levels of contextual awareness.

We extend existing reviews of the benchmarking and evaluation efforts for motion prediction (Hirakawa et al., 2018; Poiesi and Cavallaro, 2015; Quehl et al., 2017; Ridel et al., 2018) with additional datasets, probabilistic and robustness metrics, and a principled analysis of existing benchmarking practices. Furthermore, we give an up-to-date discussion of the current state of the art and conclude with recommendations for promising directions of future research.

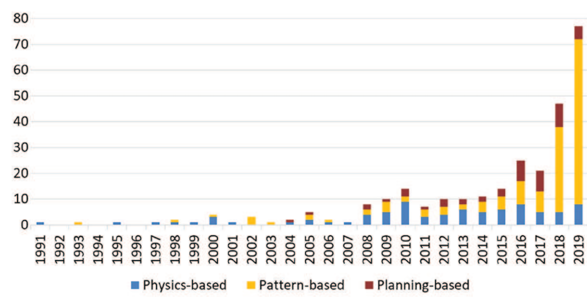

Fig. 4. Publications trends in the literature reviewed for this survey, color-coded by modeling approach.

2. Taxonomy

In this section, we describe our taxonomy to decompose the motion prediction problem based on the modeling approach and the type of contextual cues, see Figure 3 for an overview. In Sections 2.1 and 2.2 we detail the categories and give representative papers as examples of each category, and in Section 2.3 we describe the rules for classifying the methods.

2.1. Modeling approach

The motion modeling category subdivides the prediction approaches based on how they represent human motion and formulate the causes thereof. *Physics-based methods* define an explicit dynamical model based on Newton's law of motion. *Pattern-based methods* learn motion patterns from data of observed agent trajectories. *Planning-based methods* reason on motion intent of rational agents (see Figure 5). The categorization can be seen to differ also in

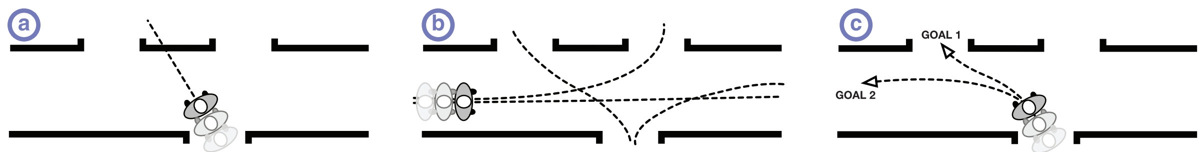

Fig. 5. Illustration of the basic working principle of the modeling approaches. (a) Physics-based methods project the motion state of the agent using explicit dynamical models based on Newton's law of motion. (b) Pattern-based methods learn prototypical trajectories from observed agent behavior to predict future motion. (c) Planning-based methods include some form of reasoning about the likely goals and compute possible paths to reach those goals. In order to incorporate internal and external stimuli that influence motion behavior, approaches can be extended to account for different contextual cues.

the level of cognition typically involved in the prediction process: physics-based methods follow a reactive sense–predict scheme, pattern-based methods follow a sense–learn–predict scheme, and planning-based methods follow a sense–reason–predict scheme in which agents reason about intentions and possible ways to the goal.

1. *Physics-based methods* (sense–predict): motion is predicted by forward simulating a set of explicitly defined dynamics equations that follow a physics-inspired model. Based on the complexity of the model, we recognize the following subclasses.
   1.1. *Single-model methods* define a single dynamical motion model (e.g., Aoude et al., 2010; Coscia et al., 2018; Elnagar, 2001; Luber et al., 2010; Pellegrini et al., 2009; Petrich et al., 2013; Yamaguchi et al., 2011; Zernetsch et al., 2016).
   1.2. *Multi-model (MM) methods* include a fixed or on-line adaptive set of multiple dynamics models and a mechanism to fuse or select the individual models (e.g., Agamennoni et al., 2012; Althoff et al., 2008a; Gindele et al., 2010; Kaempchen et al., 2004; Kooij et al., 2019; Pool et al., 2017).
2. *Pattern-based methods* (sense–learn–predict) approximate an arbitrary dynamics function from training data. These approaches are able to discover statistical behavioral patterns in the observed motion trajectories and are separated into two categories.
   2.1. *Sequential methods* learn conditional models over time and recursively apply learned transition functions for inference (e.g., Alahi et al., 2016; Aoude et al., 2011; Goldhammer et al., 2014; Keller and Gavrila, 2014; Kruse and Wahl, 1998; Kucner et al., 2017; Liao et al., 2003; Vemula et al., 2017).
   2.2. *Non-sequential methods* directly model the distribution over full trajectories without temporal factorization of the dynamics, (e.g., Bennewitz et al., 2005; Käfer et al., 2010; Keller and Gavrila, 2014; Luber et al., 2012; Tay and Laugier, 2008; Trautman and Krause, 2010; Xiao et al., 2015).
3. *Planning-based methods* (sense–reason–predict) explicitly reason about the agent's long-term motion goals and compute policies or path hypotheses that enable an agent to reach those goals. We classify the planning-based approaches into two categories.
   3.1. **Forward planning methods** make an explicit assumption regarding the optimality criteria of an agent's motion, using a pre-defined reward function (e.g., Best and Fitch, 2015; Bruce and Gordon, 2004; Galceran et al., 2015; Karasev et al., 2016; Rösmann et al., 2017; Rudenko et al., 2017; Vasquez, 2016; Xie et al., 2013; Yi et al., 2016).
   3.2. **Inverse planning methods** estimate the reward function or action model from observed trajectories using statistical learning techniques (e.g,. Chung and Huang, 2012; Huang et al., 2016; Kitani et al., 2012; Kuderer et al., 2012; Lee et al., 2017; Pfeiffer et al., 2016; Rehder et al., 2018; Shen et al., 2018; Walker et al., 2014; Ziebart et al., 2009).

Figure 4 shows the publications trends over recent years, color-coded by modeling approach. The number of related works is strongly increasing during the last two years in particular for the pattern-based methods.

2.2. Contextual cues

We define contextual cues to be all relevant internal and external stimuli that influence motion behavior and categorize them based on their relation to the target agent, other agents in the scene and properties of the static environment, see Figures 6 and 7.

1. Cues of the *target agent* include:
   1.1. *motion state* such as position and possibly velocity (e.g., Bennewitz et al., 2005; Bera et al., 2016; Elfring et al., 2014; Ferrer and Sanfeliu, 2014; Karasev et al., 2016; Kitani et al., 2012; Kooij et al., 2019; Kucner et al., 2017; Kuderer et al., 2012; Pellegrini et al., 2009; Trautman and Krause, 2010; Ziebart et al., 2009);

   1.2. *articulated pose* such as head orientation (e.g., Hasan et al., 2018; Kooij et al., 2019, 2014; Roth et al., 2016; Unhelkar et al., 2015) or full-body pose (Mínguez et al., 2018; Quintero et al., 2014);
   1.3. *semantic attributes* such as the age and gender (Ma et al., 2017), personality (Bera et al., 2017), and awareness of the robot's presence (Kooij et al., 2019; Oli et al., 2013).

2. With respect to the *dynamic environment* we distinguish:

   2.1. *unaware methods*, which compute motion predictions for the target agent not considering the presence of other agents (e.g., Bennewitz et al., 2005; Elnagar, 2001; Elnagar and Gupta, 1998; Kim et al., 2011; Kucner et al., 2013; Thompson et al., 2009; Wang et al., 2016; Zhu, 1991);
   2.2. *individual-aware methods*, which account for the presence of other agents (e.g., Alahi et al., 2016; Elfring et al., 2014; Ferrer and Sanfeliu, 2014; Kooij et al., 2019; Kuderer et al., 2012; Luber et al., 2010; Trautman and Krause, 2010; Vemula et al., 2017);
   2.3. *group-aware methods*, which account for the presence of other agents as well as social grouping information; this allows agents to be considered in groups, formations, or convoys that move differently than independent agents (e.g., Karamouzas and Overmars, 2012; Pellegrini et al., 2010; Qiu and Hu, 2010; Robicquet et al., 2016; Seitz et al., 2012; Singh et al., 2009; Yamaguchi et al., 2011).

3. With respect to the *static environment* we distinguish:

   3.1. *unaware methods*, which assume an open-space environment (e.g., Bennewitz et al., 2002; Ellis et al., 2009; Ferguson et al., 2015; Foka and Trahanias, 2010; Jacobs et al., 2017; Kruse and Wahl, 1998; Luber et al., 2012; Schneider and Gavrila, 2013; Unhelkar et al., 2015; Vasquez et al., 2008);
   3.2. *obstacle-aware methods*, which account for the presence of individual static obstacles (e.g., Alahi et al., 2016; Althoff et al., 2008b; Bera et al., 2016; Elfring et al., 2014; Ferrer and Sanfeliu, 2014; Rehder and Klöden, 2015; Trautman and Krause, 2010; Vemula et al., 2017);
   3.3. *map-aware methods*, which account for environment geometry and topology (e.g., Chen et al., 2017; Chung and Huang, 2010, 2012; Gong et al., 2011; Henry et al., 2010; Ikeda et al., 2012; Kooij et al., 2019; Liao et al., 2003; Pfeiffer et al., 2016; Pool et al., 2017; Rösmann et al., 2017; Rudenko et al., 2017, 2018b; Vasquez, 2016; Yen et al., 2008; Ziebart et al., 2009);
   3.4. *semantics-aware methods*, which additionally account for environment semantics or affordances such as no-go zones, crosswalks, sidewalks, or traffic lights (e.g., Ballan et al., 2016; Coscia et al., 2018; Karasev et al., 2016; Kitani et al., 2012; Kuhnt et al., 2016; Lee et al., 2017; Ma et al., 2017; Rehder et al., 2018; Zheng et al., 2016).

In the following Sections 3, 4, and 5 we survey the different classes of the motion model category. We detail contextual cues categories in Section 6. In each section, we discuss methods in the order of increasing complexity, considering inheritance of ideas and grouped by the similarity of the motion modeling techniques.

2.3. Classification rules

Some of the surveyed papers may not fall univocally into a single class of our taxonomy, especially those using a mixture of different approaches, e.g., the work by Bennewitz et al. (2005) that combines a non-sequential clustering approach with sequential hidden Markov model (HMM) inference. For those borderline cases, we adopt the following rules.

(*i*) We classify methods primarily in the category that best describes the modeling approach over the inference method, e.g., for Bennewitz et al. (2005) we give more weight to the clustering technique used for modeling the usual human motion behavior.

(*ii*) Some approaches add sub-components from other categories in their main modeling approach, e.g., planning-based approaches using physics-based transition functions (Rudenko et al., 2018a; van Den Berg et al., 2008), physics-based methods tuned with learned parameters (Ferrer and Sanfeliu, 2014), planning-based approaches using IRL to recover the hidden reward function of human behaviors (Kitani et al., 2012; Ziebart et al., 2009). We classify such approaches based on their main modeling method.

(*iii*) Methods that use behavior cloning (imitation of human behaviors with supervised learning techniques), i.e., learn/recover the motion model directly from data, are classified as pattern-based approaches (Schmerling et al., 2018; Zheng et al., 2016). In contrast to that, imitation learning techniques that reason on policies (e.g., using generative adversarial imitation learning (GAIL) (Li et al., 2017)) are classified as planning-based methods.

Furthermore, a single work is categorized into three contextual cues' classes with respect to its perception of the target agent, static and dynamic contextual cues.

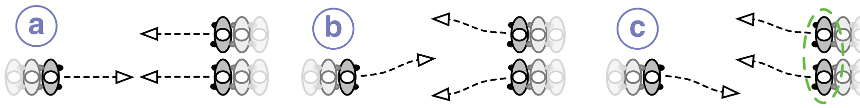

Fig. 6. Dynamic environment cues: (a) unaware, (b) individual-aware, (c) group-aware (accounting for social grouping cues, in green).

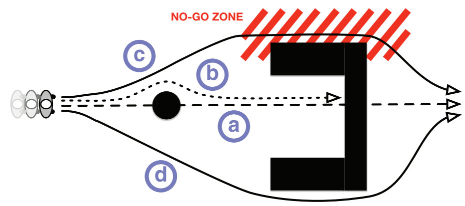

Fig. 7. Static environment cues: (a) unaware (ignoring any static objects, dashed line), (b) obstacle-aware (accounting for unmodeled obstacles, dotted line), (c) map-aware (accounting for a topometric environment model avoiding local minima, solid line), (d) semantics-aware (solid line).

3. Physics-based approaches

Physics-based models generate future human motion considering a hand-crafted, explicit dynamical model f based on Newton's laws of motion. A common form for f is $\dot{\mathbf{s}}_t = f(\mathbf{s}_t, \mathbf{u}_t, t) + w_t$ where $\mathbf{u}_t$ is the (unknown) control input and w_t the process noise. In fact, motion prediction can be seen as inferring $\mathbf{s}_t$ and $\mathbf{u}_t$ from various estimated or observed cues.

A large variety of physics-based models have been developed in the target tracking and automatic control communities to describe motion of dynamic objects in ground, marine, airborne, or space applications, typically used as building blocks of a recursive Bayesian filter or multiple-model algorithm. These models differ in the type of motion they describe such as maneuvering or non-maneuvering motion in two or three dimensions, and in the complexity of the target's kinematic or dynamic model and the complexity of the noise model. See Li and Jilkov (2003, 2010) for a survey on physics-based motion models for target tracking.

We subdivide physics-based models into (1) *single-model approaches* that rely on a single dynamical model f and (2) *MM approaches* that involve several modes of dynamics (see Figure 8).

3.1. Single-model approaches

3.1.1. Early works and basic models. Many approaches to human motion prediction represent the motion state of target agents as position, velocity, and acceleration and use different physics-based models for prediction. Among the simplest ones are kinematic models without considering forces that govern the motion. Popular examples include the constant velocity model (CV) that assumes piecewise constant velocity with white noise acceleration, the constant acceleration model (CA) that assumes piecewise constant acceleration with white noise jerk, the coordinated turn model (CT) that assumes constant turn rate and speed with white noise linear and white noise turn acceleration or the more general curvilinear motion model by Best and Norton (1997). The bicycle model is often used as an approximation to model the vehicle dynamics (see, e.g., Schubert et al., 2008).

A large number of works across all application domains rely on kinematic models for their simplicity and acceptable performance under mild conditions such as tracking with little motion uncertainty and short prediction horizons. Examples include Møgelmose et al. (2015) for hazard inference from linear motion predictions of pedestrians or Elnagar (2001) for Kalman filter-based (KF) prediction of dynamic obstacles using a constant acceleration model. Barth and Franke (2008) used the coordinated turn model for one-step-ahead prediction in an extended Kalman filter (EKF) to track oncoming vehicles from point clouds generated by an in-car stereo camera. Batz et al. (2009) used a variant of the coordinated turn model for one-step motion prediction of vehicles within an unscented Kalman filter to detect dangerous situations based on predicted mutual distances between vehicles.

Dynamic models account for forces which, following Newton's laws, are the key descriptor of motion. Such models can become complex when they describe the physics of wheels, gearboxes, engines, or friction effects. In addition to their complexity, forces that govern the motion of other agents are not directly observable from sensory data. This makes dynamic models more challenging for motion prediction. Zernetsch et al. (2016) used a dynamic model for trajectory prediction of cyclists that contains the driving force and the resistance forces from acceleration, inclination, rolling, and air. The authors showed experimentally that long-term predictions up to 2.5 se ahead are geometrically more accurate when compared with a standard CV model.

Autoregressive models (ARMs) that, unlike first-order Markov models, account for the history of states have also been used for motion prediction. Elnagar and Gupta (1998) employed a third-order ARM to predict the next position and orientation of moving obstacles using maximum-likelihood estimation of the ARM parameters. Cai et al.

(2006) used a second-order ARM for single step motion prediction within a particle filter for visual target tracking of hockey players. The early work by Zhu (1991) used an autoregressive moving average model as transition function of a HMM to predict occupancy probabilities of moving obstacles over multiple time steps with applications to predictive planning.

Physics-based models are used for motion prediction by recursively applying the dynamics model f to the current state of the target agent. So far, with the exception of Zhu (1991), the works described above make only one-step-ahead predictions and ignore contextual cues from the environment. To account for context, the dynamics model f can be extended by additional forces, model parameters, or state constraints as discussed hereafter.

3.1.2. Models with map-based contextual cues. A number of approaches extend physics-based models to account for information from a map, particularly for the task of tracking ground vehicles on roads. The methods developed to this end differ in how road constraints are derived and incorporated into the state estimation problem, see the survey by Simon (2010). Yang and Blasch (2008), for example, used a regular KF and project the unconstrained state estimate onto the constrained surface for tracking on-road ground vehicles with a surveillance radar. Yang et al. (2005) used the technique to reduce the system model parametrization to the constrained surface. They reduced vehicle motion to a 1D curvilinear road representation for filtering. Batkovic et al. (2018) predicted pedestrian motion along a graph with straight line edges centered on sidewalks and crosswalks. Using a unicycle model and a control approach to keep the predictions along the edges, they evaluated long-term predictions up to 10 s ahead. When there are several possible turns at a node, i.e., at bifurcations, predictions are propagated along all outgoing edges. Another class of techniques uses the road information as pseudo-measurements, pursued, e.g., by Petrich et al. (2013) who used a kinematic bicycle model for f and pseudo-measurements from the centerlines of lanes to predict future vehicle trajectories several seconds ahead. When there are several possible turns, e.g., at intersections, the approach generates new motion hypothesis for each relevant lane by using an EKF.

When agents move freely, e.g., do not comply with road constraints, we need different ways to represent free space and account for map information. To this end, several authors proposed grid-based (Coscia et al., 2018; Luber et al., 2011; Rehder and Klöden, 2015) and more general graph-based space discretizations (Aoude et al., 2010; Koschi et al., 2018). Luber et al. (2011) used 2D laser data to track people from a mobile robot and learn a so-called spatial affordance map, a grid-based spatial Poisson process from which a walkable area map of the environment can be derived. They predict future trajectories of people during lengthy occlusion events using an auxiliary PF with look-ahead particles obtained by forward-simulation of the curvilinear motion model proposed by Best and Norton (1997). In this way, long-term predictions (up to 50 steps ahead) stay focused on high-probability regions with the result of improved tracking performance. Rehder and Klöden (2015) also chose a regular grid to represent the belief about pedestrian locations in a linear road scenario. They proposed a variant of a Bayesian histogram filter to achieve map-aware predictions 3 seconds ahead by combining forward propagation of an unicycle pedestrian model from the start and in backward direction from the goal with prior place-dependent knowledge of motion learned from previously observed trajectories. Similarly, Coscia et al. (2018) used polars grids, centered at the currently predicted agent position to represent four different local influences: a CV motion model, prior motion knowledge learned from data, semantic map annotations such as "road" or "grass," and direction to goal. The next velocity is then obtained from the normalized product of the four polar distributions and forward propagated for long-term prediction of pedestrians and cyclists in urban scenarios. Like Rehder and Klöden (2015), no planning is involved and the learned prior knowledge is place-dependent. Koschi et al. (2018) exploited information on road segments connectivity and semantic regions to compute reachability-based predictions of pedestrians, similarly to Rehder and Klöden (2015). The authors formalized several relevant traffic rules, e.g., pedestrian crossing permission on the green light, as additional motion constraints. Aoude et al. (2010) grew a tree of future trajectories for each target agent using a closed-loop rapidly exploring random trees (RRT) algorithm that samples the controls of a bicycle motion model (Kuwata et al., 2009) avoiding obstacles in the map. Based on agent's recognized intentions using a support vector machine (SVM) classifier and features from observed trajectories, they bias the tree growth towards areas that are more likely for the agent to enter and determine the best evasive maneuver for the ego-vehicle to minimize threat at intersection scenarios. A reachibility-based model, such as Rehder and Klöden (2015), Koschi et al. (2018), and Aoude et al. (2010), is illustrated in Figure 8(b).

So far, we discussed extensions to physics-based motion models that embed different types of map information. All those works, however, consider only a single target agent and neglect local interactions between multiple agents. Hereafter, we discuss methods that add social situation awareness, predicting several target agents jointly.

3.1.3. Models with dynamic environment cues. There are several ways to incorporate local agent interaction models into physics-based approaches for prediction, one popular example being the social force (SF) model by Helbing and Molnar (1995), see Figure 8 (c). Developed for the purpose of crowd analysis and egress research, the model superimposes attractive forces from a goal with repulsive forces from other agents and obstacles. Several works extend

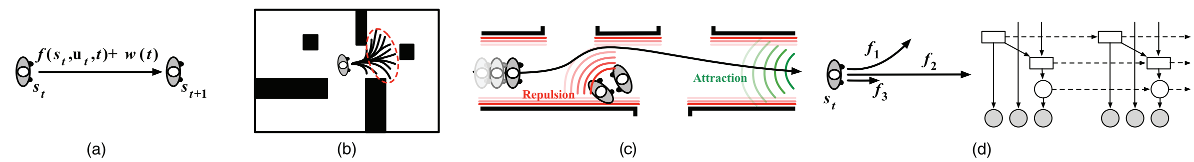

Fig. 8. Examples of the physics-based approaches: (a) a method with a single dynamical model, (b) a reachability-based method, which accounts for all possible transitions from the given motion state, (c) an attraction–repulsion approach, which accounts for dynamic environment cues, (d) a MM method with several modes of dynamics and the DBN switching mechanism.

the dynamics model f to include social forces, e.g., for improved short-term prediction for pedestrian tracking in 2D laser data (Luber et al., 2010) or image data (Pellegrini et al., 2009).

Elfring et al. (2014) combined the HMM-based goal estimation method introduced by Vasquez et al. (2008) with the basic SF-based human motion prediction by Luber et al. (2010). For intention estimation, the observed people trajectories are summarized in a sparse topological map of the environment. Each node of the map encodes a state–destination pair, and the goal inference using the observed trajectory is carried out in a maximum-likelihood manner. Ferrer and Sanfeliu (2014) estimated the interaction parameters of the SF for each two people in the scene individually. For this purpose, several *behaviors* (i.e., sets of SF parameters) are learned offline, and the observed interaction between any two people is associated to the closest "behavior." The approach by Oli et al. (2013) defined the robot operating in social spaces as an interacting agent, affected by the social forces. Each human was flagged as either aware or unaware of the robot, which defined the repulsive force the robot exerts on that person. Such awareness was inferred using visual cues (gaze direction and past trajectory).

In order to achieve more realistic behaviors, several extensions to the social force model have been proposed. Yan et al. (2014) presented a model that embeds social relationships in the linear combination of predefined basic social effects (attraction, repulsion, and non-interaction). The motion predictor maintains several hypothesis over the social modes, in which the pedestrians are involved. Predictive collision avoidance behavior of the SF agents is introduced by Karamouzas et al. (2009) and Zanlungo et al. (2011). In particular, Karamouzas et al. (2009) modeled each agent to adapt their route as early as possible, trying to minimize the amount of interactions with others and the energy required to solve these interactions. To this end an evasion force, that depends on the predicted point of collision and the distance to it, is applied to each agent. Updates to the SF model to consider also group motion have been proposed by Moussaïd et al. (2010) and Farina et al. (2017).

Other agent interaction models, not based on the social forces, for example for road vehicles, have also been used. An interactive kinematic motion model for vehicles on a single lane has been proposed by Treiber et al. (2000) to predict the longitudinal motion of a target vehicle in the presence of preceding vehicles. The model, called the intelligent driver model (IDM), was used, e.g., by Liebner et al. (2013) for driver intent inference at urban intersections. Hoermann et al. (2017) learned the driving style of preceding vehicles by on-line estimating the IDM parameters using particle filtering and near- and far-range radar observations. Prediction of longitudinal motion of preceding vehicles, in the experiments up to 10 seconds ahead, is then obtained by forward propagation of the model.

Several approaches exploit the *reciprocal velocity obstacles* (RVO) model (van den Berg et al., 2008) for jointly predicting human motions. Kim et al. (2015) used the ensemble KF technique together with the expectation–maximization (EM) algorithm to estimate and improve the human motion model (i.e., RVO parameters). Bera et al. (2016) proposed a method that dynamically estimates parameters of the RVO function for each pedestrian, moving in a crowd, namely current and preferred velocities per agent and global motion characteristics such as entry points and movement features. A follow-up work (Bera et al., 2017) also introduced online estimation of personality traits. Each pedestrian's behavior is characterized as a weighted combination of six personality traits (aggressive, assertive, shy, active, tense, and impulsive) based on the observations, thus defining parameters of the RVO model for this person.

Other approaches instead compute joint motion predictions based on the time of possible collision between pairs of agents. Paris et al. (2007) proposed a method for modeling predictive collision avoidance behavior in simulated scenarios. For each pedestrian current velocities of their neighbors are extrapolated in the 3D (x, y, t) space, and all actions that result in collision with dynamic and static obstacles are excluded. A similar problem was addressed by Pettré et al. (2009), who evaluated real people trajectories in an interactive experiment and designed a predictive collision avoidance approach, capable of reproducing realistic joint maneuvers, such as giving way and passing first.

Other methods propose to compute joint motion prediction based on the expected point of closest approach between pedestrians. The first such approach was proposed by Pellegrini et al. (2009) called *linear trajectory avoidance* (LTA): the method first computes the expected point of closest approach between different agents, and then uses it as driving force to perform avoidance between the agents.

Based on the LTA, Yamaguchi et al. (2011) formulated a human motion prediction approach as an energy minimization problem. The energy function considers different properties of people motion: damping, speed, direction, attraction, being in a group, avoiding collisions. The approach of Yamaguchi et al. was further improved by Robicquet et al. (2016) by considering several different sets of the energy functional parameters, learned from the training data. Each set of parameters represents a distinct behavior (navigation style of the agent).

Local interaction modeling methods, as well as approaches for predicting motion in crowds, usually benefit from detecting and considering groups of people who walk together. For example, Pellegrini et al. (2010) proposed an approach to model joint trajectories of people, taking group relations into account. The proposed framework operates in two steps: first, it generates possible trajectory hypotheses for each person, then it selects the best hypothesis that maximize a likelihood function, taking into account social factors, while at the same time estimating group membership. People and relations are modeled with conditional random fields (CRFs). Choi and Savarese (2010) proposed an interaction model that incorporates linear motion assumption, repulsion of nearby people and group coherence via synchronization of velocities. Further group motion models (e.g., Karamouzas and Overmars, 2012; Qiu and Hu, 2010; Seitz et al., 2012; Singh et al., 2009), developed in the simulation and visualization communities, typically addresses the groups cohesion with additional forces to attract members to each other, assigning leader's and follower's roles or imposing certain group formation.

A recent reachability-based pedestrian occupancy prediction method, presented by Zechel et al. (2019), accounts both for dynamic objects and semantics of the static environment. The authors first used a physical model to determine reachable locations of a person, and then reduced the area based on the intersections with static environment and presence probabilities of other dynamic agents. Similarly, Luo and Cai (2019) computed future agents predictions based on an optimization approach that handles physical constraints, i.e., kinematics and geometry of the agents, and behavioral constraints, i.e., intention, attention, and responsibility.

3.2. Multi-model approaches

Complex agent motion is poorly described by a single dynamical model f. Although the incorporation of map information and influences from multiple agents render such approaches more flexible, they remain inherently limited. A common approach to modeling general motion of maneuvering targets is the definition and fusion of different prototypical motion modes, each described by a different dynamic regime f. Modes may be linear movements, turn maneuvers, or sudden accelerations, that, over time, form sequences able to describe complex motion behavior. Since the motion modes of other agents are not directly observable, we need techniques to represent and reason about motion mode uncertainty. The primary approach to this end are MM methods (Li and Jilkov, 2005) and hybrid estimation (Hofbaur and Williams, 2004). MM methods maintain a hybrid system state $\xi = (\mathbf{x}, s)$ that augments the continuous valued $\mathbf{x}$ by a discrete-valued modal state s. Following Li and Jilkov (2005), MM methods generally consist of four elements: a fixed or on-line adaptive model set, a strategy to deal with the discrete-valued uncertainties (e.g., model sequences under a Markov or semi-Markov assumption), a recursive estimation scheme to deal with the continuous valued components conditioned on the model, and a mechanism to generate the overall best estimate from a fusion or selection of the individual filters. For prediction, MM methods are used in several ways, to represent more complex motion, to incorporate context information from other agents and context information from the map. A naive MM approach, presented by Pool et al. (2017), predicted future motion of cyclists using a uniform mixture of five linear dynamic systems (LDSs) dynamics-based motion strategies: go on straight, turn 45° or 90° left or right. The probability of each strategy is set to zero if the predicted path does not comply with the road topology in the place of prediction.

The interactive multiple model filter (IMM) is a widely used inference technique applied on MM models with numerous applications in tracking (Mazor et al., 1998) and predictions. For instance, Kaempchen et al. (2004) proposed a method for future vehicle states estimation that switches between constant acceleration and simplified bicycle dynamical models. Uncertainty in the next transition is explicitly modeled with Gaussian noise. Schneider and Gavrila (2013) introduce an IMM for pedestrian trajectory prediction that combines several basic motion models (constant velocity, constant acceleration, and constant turn). In addition, Schulz and Stiefelhagen (2015) proposed a method for predicting the future path of a pedestrian using an IMM framework with constant velocity, constant position, and coordinated turn models. In this work, model transitions are controlled by an intention recognition system based on latent-dynamic conditional random fields: based on the features of the person's dynamics (position and velocity) and situational awareness (head orientation), intention is classified as crossing, stopping, or going in the same direction. Joint vehicle trajectory estimation also using IMMs is considered by Kuhnt et al. (2015, 2016) in a method which adopts pre-defined environment geometry to estimate possible routes of each individual vehicle. Contextual interaction constraints are embedded in a Bayesian network that estimates the evolution of the traffic situation.

Other examples of IMMs techniques are variable-structure IMMs for ground vehicles (Kirubarajan et al., 2000; Noe and Collins, 2000; Pannetier et al., 2005; Shea et al., 2000) to account for road constraints. In a recent work, Xie et al. (2018) combined a kinematics-based constant turn rate and acceleration model with IMM-based lane keeping and changing maneuvers mixing. The method is aware of road geometry and produces results for a varying prediction horizon.

An alternative approach to hybrid estimation problems are dynamic Bayesian networks (DBNs), which inherit the broad variety of modeling schemes and large corpus of exact and approximate inference and learning techniques from probabilistic graphical models (Koller et al., 2009). An example of a DBN-based MM approach is given in Figure 8(d). The seminal work of Pentland and Liu (1999) introduced an approach to model human behaviors by coupling a set of dynamic systems (i.e., a bank of KFs) with an HMM, which is a special case of the DBNs. The authors introduce a dynamic Markov system that infers human future behaviors, a set of macro-actions described by a set of KFs, based on measured dynamic quantities (i.e., acceleration, torque). The approach was used to accurately categorize human driving actions. Agamennoni et al. (2012) jointly modeled the agent dynamics and situational context using a DBN. The vehicular dynamics is described by a bicycle model whereas the context is defined by a weighted feature function to account, e.g., for closeness between agents or place-dependent information from a map. The model resembles a switched Bayesian filter, but considers a more general conditioning of the switch transitions and the case of multiple agents. The authors applied the model for the task of long-term multi-vehicle trajectory prediction of mining vehicles, useful for instance during GPS outages. Kooij et al. (2014) proposed a context-aware path prediction method for pedestrians intending to laterally cross a street, that makes use of switching linear dynamical systems (SLDSs) to model maneuvering pedestrians that alternate between motion models (e.g., walking straight, stopping). The approach adopts a DBN to infer the next pedestrian movements based on the SLDS model. The latent (context) variables relate to pedestrian awareness of an oncoming vehicle (head orientation), the distance to the curbside and the situation criticality. Kooij et al. (2019) extended this work to cover a cyclist turning scenario. In another extension of Kooij et al. (2014), Roth et al. (2016) used a second context-based SLDS to model the “braking” and “driving” behaviors of the ego-vehicle. The two SLDS sub-graphs for modeling pedestrian and vehicle paths are combined into a joint DBN, where the situation criticality latent state is shared. Gu et al. (2016) proposed a DBN-based motion model with a particle filter inference to estimate future position, velocity, and crossing intention of a pedestrian. During inference the approach considers standing, walking, and running motion modes of pedestrians. Gindele et al. (2010) jointly modeled future trajectories of vehicles with a DBN, describing the local context of the interaction between multiple drivers with a set of numerical features. These features were used to classify the current situation of each driver and reason on available behaviors, such as “follow,” “sheer in,” or “overtake,” represented as Bézier curves. Blaiotta (2019) also proposed a DBN for pedestrian prediction with two motion modes (walking and standing), contextual awareness flag for the oncoming vehicle and social force-based motion dynamics for pedestrians.

Techniques derived by the stochastic reachability analysis theory (Althoff, 2010) form another class of hybrid approaches to compute human motion prediction. In general, those methods model agents as hybrid systems (with multiple modes) and infer agents’ future motions by computing stochastic reachable sets. The approach by Althoff et al. (2008b) generates the stochastic reachable sets for interacting traffic participants using Markov chains, where each chain approximates the behavior of a single agent. Each vehicle has its own dynamics with many modes (e.g., acceleration, deceleration, standstill, speed limit), and its goal is assumed to be known. Althoff et al. (2013) further extended Althoff et al. (2008b) with the over-approximative estimation of the occupancy sets. The method is particularly framed for hybrid dynamics (mixed discrete and continuous) where computing the exact reachability sets could be computationally unfeasible. To overcome this issue, the method proposes to intersect different occupancy sets for different abstractions of the dynamical model. The work by Bansal et al. (2019) also used a reachability approach for solving the prediction problem for MM systems. The approach rather than using a probability distribution over human next actions, it uses a deterministic set of allowable human actions. This reduces the complexity of the predictor and allows for an easy certification process.

4. Pattern-based approaches

In contrast to the physics-based approaches that use explicitly defined, parametrized functions of motion dynamics, pattern-based approaches learn the latter from data, following the *sense–learn–predict* paradigm. These methods learn human motion behaviors by fitting different function approximators (i.e., neural networks, HMMs, Gaussian processes (GPs)) to data. Many of those methods were introduced by the machine learning and computer vision communities (i.e., for behavior cloning and video surveillance applications), and later applied in robotics and autonomous navigation settings.

In our taxonomy we classify pattern-based approaches into two categories, based on the type of function approximator used.

(1) *Sequential methods* typically learn conditional models, where it is assumed that the state (e.g., position, velocity) at one time instance is conditionally dependent on some sufficient statistic of the full history of past states. Many of the proposed methods are Markov models, where an Nth-order Markov model assumes that a limited state history of N time steps is a sufficient representation of the entire state history. Similarly to many physics-based approaches, sequential methods aim to learn a one-step predictor $\mathbf{s}_{t+1} = f(\mathbf{s}_{t-n:t})$, where the state $\mathbf{s}_{t+1}$ is the one step prediction

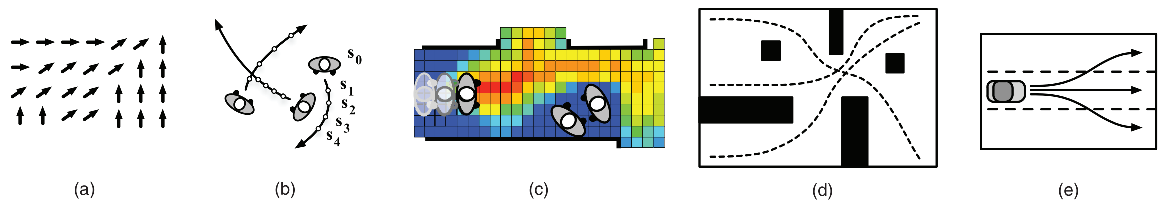

Fig. 9. Examples of the pattern-based approaches: (a) grid-based local transitions learning method, (b) sequential location-independent transition model, which accounts for cues from the dynamic environment, (c) higher-order sequential Markov model, (d) clustering of full trajectories, (e) location-independent method, which learns long-term transition sequences, i.e., maneuvers.

and the sequence of states $\mathbf{s}_{t-n:t}$ is the sufficient statistic of the history. In order to predict a sequence of state transitions (i.e., a trajectory), consecutive one-step predictions are made to compose a single long-term trajectory.

(2) *Non-sequential methods* directly model the distribution over full trajectories without imposing a factorization of the dynamics (i.e., Markov assumption) as with sequential models.

4.1. Sequential models

Sequential models are built on the assumption that the motion of intelligent agents can be described with causally conditional models over time. Similarly to the physics-based methods, transition function of sequential models has Markovian property, i.e., information on the future motion is confined in the current state of the agent. Differently, the function, often non-parametric (e.g., GPs, vector fields), is learned from statistical observations, and its parameters cannot be directly interpreted as for many of the physics-based methods.

4.1.1. Local transition patterns, Learning local motion patterns, such as probabilities of transitions between cells on a grid map (Figure 9(a)), is a simple, commonly used technique for making sequential predictions (Ballan et al., 2016; Kruse and Wahl, 1998; Kucner et al., 2013; Molina et al., 2018; Tadokoro et al., 1993; Thompson et al., 2009; Wang et al., 2015, 2016).

Early examples of local motion patterns include the works of Tadokoro et al. (1993) and Kruse and Wahl (1998). Kruse and Wahl (1998) built two transition models: a stochastic grid where usual motion patterns of dynamic obstacles are stored, and stochastic trajectory prediction modeled with Poisson processes. Tadokoro et al. (1993) included empirical biases to account for context features of the cells in the regions where the observations are sparse, e.g., increasing the probability to move away from the wall, stop near a bookshelf or decrease walking speed at the crossing. More recently, Thompson et al. (2009) expanded the local motion patterns model by accounting for further transitions for several steps into the future. Their method maps the motion state of the person to a series of local patches, describing where the person might be in the future. In addition to the current motion state, the learned patterns are also conditioned on the final goal or the topological sub-goal in the environment. Wang et al. (2015) modeled local transition probabilities with an input–output HMM. Transition in each cell is conditioned both on the direction of cell entrance and the global starting point of the person's movement. Jacobs et al. (2017) used nonlinear estimation of pedestrian dynamics with the learned vector fields to improve the linear velocity projection model. Ballan et al. (2016) proposed a DBN method to predict not-interacting human motion based on statistical properties of human behavior. To this end a transferable navigation grid-map is learned. It encodes functional properties of the environment (i.e., direction and speed of the targets, crossing frequency for each patch, identification of routing points). Molina et al. (2018) addressed periodic temporal variations in the learned transition patterns, e.g., based on the time of the day.

In contrast to the discrete transition patterns discussed so far, several authors modeled the transition dynamics as a continuous function of the agent's motion state, using GPs and their mixtures (Ellis et al., 2009; Ferguson et al., 2015; Joseph et al., 2011; Kucner et al., 2017). Ellis et al. (2009) modeled trajectory data in the observed environment by regressing relative motion against current position. Predictions were generated using a sequential Monte Carlo sampling method. Joseph et al. (2011) modeled the multi-modal mobility patterns as a mixture of GPs with a Dirichlet process prior over mixture weights. Ferguson et al. (2015) further extended the work of Joseph et al. (2011) by including a change-point detection and clustering algorithm that enables quick detection of changes in intent and on-line learning of motion patterns not seen in prior training data. Kucner et al. (2017) model multimodal distributions with a Gaussian mixture model (GMM) in the joint velocity–orientation space.

Apart from the commonly used grid cells, local transition patterns can be learned using a higher-level abstraction of the workspace, such as a graph of sub-goals or transition points (Han et al., 2019; Ikeda et al., 2012), map of connected position–velocity points Kalayeh et al. (2015), Voronoi diagram (Liao et al., 2003), instantaneous topological map (ITM) (Vasquez et al., 2009), and semantic-aware ITM (Vasishta et al., 2018). More flexible representation of the workspace topology is achieved this way. Combining the merits of local and global motion patterns (i.e., sequential and non-sequential models), Chen et al. (2016) modeled trajectories in the environment with a set of overcomplete basis vectors. The method breaks down trajectories into a small number of representative partial motion patterns, where each partial pattern consists of a series of local transitions. A follow-up work by Habibi et al. (2018) incorporated semantic features from the environment (relative distance to curbside and the traffic lights signals) in the learning process, improving prediction accuracy and generalization to similar environments. Han et al. (2019) proposed a method to explicitly learn transition points between the local patterns.

4.1.2. Location-independent behavioral patterns. Unlike the local transition patterns, which are learned and applied for prediction only in a particular environment, *location-independent* patterns are used for predicting transitions of an agent in the general free space (Aoude et al., 2011; Foka and Trahanias, 2002; Quintero et al., 2014; Shalev-Shwartz et al., 2016; Tran and Firl, 2014) (see Figure 9(b)).

Several authors (e.g., Foka and Trahanias, 2002; Shalev-Shwartz et al., 2016) used location-invariant one-step prediction as a part of collision avoidance framework using neural networks. Aoude et al. (2011) extended their physics-based approach (Aoude et al., 2010) by introducing location-independent GP-based motion patterns that guide the RRT-Reach to grow probabilistically weighted feasible paths of the surrounding vehicles. Tran and Firl (2014) modeled location-independent motion patterns of vehicles by applying spatial normalization to the trajectories in the learning set. Cartesian coordinates are turned into the relative coordinate system of the road intersection, based on the topology of the lanes.

Keller and Gavrila (2014) used optical flow features derived from a detected pedestrian bounding box to predict future motion. Quintero et al. (2014) instead extracted full-body articulated pose. In both works, body motion dynamics for walking and stopping are learned using GPs with dynamic model (GPDM) in a compact low-dimensional latent space. Mínguez et al. (2018) extended the work of Quintero et al. (2014) by considering standing and starting activities as well. A first-order HMM is used to model the transition between the activities.

Several location-independent methods learned socially aware models of local interactions (Antonini et al., 2006; Vemula et al., 2017). Antonini et al. (2006) adapted the discrete choice model from econometrics studies to predict local transitions of individuals, given the intended direction, current velocity, locations of obstacles, and other people nearby. Vemula et al. (2017) reformulated the non-sequential joint human motion prediction approach by Trautman and Krause (2010), discussed in Section 4.2, as sequential inference with GPs. They modeled the local motion of each agent conditioned on relative positions of other people in the surroundings and the person's goal.

4.1.3. Complex long-term dependencies. Several recent sequential methods use neural networks for time series prediction, i.e., assuming a higher-order Markov property (Alahi et al., 2016; Bartoli et al., 2018; Goldhammer et al., 2014; Jain et al., 2016; Schmerling et al., 2018; Sumpter and Bulpitt, 2000; Sun et al., 2018; Varshneya and Srinivasaraghavan, 2017; Vemula et al., 2018; Zheng et al., 2016), see Figure 9(c). Such time series-based models are making a natural transition between the first-order Markovian methods (e.g., local transition patterns) and non-sequential techniques (e.g., clustering-based). An early method, presented by Sumpter and Bulpitt (2000), learned long-term spatiotemporal motion patterns from visual input in a known environment. The simple neural network architecture, based on natural language processing networks, quantizes partial trajectories in location/shape-space: the symbol network categorizes the object shape and locations at any time, and the context network categorizes the order in which they appear. Goldhammer et al. (2014) learned usual human motion patterns using an artificial neural network (ANN) with the multilayer perceptron architecture. This method was adapted to predict motion of cyclists by Zernetsch et al. (2016).

Recurrent neural networks (RNNs) for sequence learning, and long short-term memory (LSTM) networks in particular, have recently become a widely popular modeling approach for predicting human (Alahi et al., 2016; Bartoli et al., 2018; Sadeghian et al., 2019; Saleh et al., 2018b; Sun et al., 2018; Varshneya and Srinivasaraghavan, 2017; Vemula et al., 2018), vehicle (Altché and de La Fortelle, 2017; Ding et al., 2019; Kim et al., 2017; Park et al., 2018), and cyclist (Pool et al., 2019) motion. Alahi et al. (2016) was the first to propose a Social-LSTM model to predict joint trajectories in continuous spaces. Each person is modeled by an individual LSTM. Since humans are influenced by nearby people, LSTMs are connected in the social pooling system, sharing information from the hidden state of the LSTMs with the neighboring pedestrians. The work of Bartoli et al. (2018) extended the Social-LSTM, explicitly modeling human–space interactions by defining a "context-aware" pooling layer, which considers the static objects in the neighborhood of a person. Varshneya and Srinivasaraghavan (2017) used a spatial matching network (SMN), first introduced by Huang et al. (2016) (discussed in Section 5.2), that models the spatial context of the surrounding environment, predicting the probability of the subject stepping on a particular patch. Sun et al. (2018) used LSTM to learn environment- and time-specific human activity patterns in the target environment from long-term observations, i.e., covering several weeks. The state of the person is extended to include contextual information, i.e., the time of the day when the person is observed. Pfeiffer et al. (2018) coupled obstacle-awareness with an efficient representation of the surrounding dynamic agents using a 1D vector in polar angle space. Bisagno et al. (2018) added group coherence information in the social pooling layer. Saleh et al. predicted trajectories of pedestrians (Saleh et al., 2018b) and cyclists (Saleh et al., 2018a), adapting the LSTM architecture for the perspective of a moving vehicle. Numerous other implementations of the LSTM-based predictors offer various improvements, such as increased generalizability to new and crowded environments (Shi et al., 2019; Xue et al., 2019), considering the immediate (Zhang et al., 2019) or long-term (Xue et al., 2017) intention of the agents, augmenting the state of the person with the head pose (Hasan et al., 2018) or adding a better pooling mechanism with relative importance of each person in the vicinity of the target agent (Fernando et al., 2018; Pei et al., 2019; Xu et al., 2018). Huynh and Alaghband (2019) applied LSTM-based trajectory prediction in combination with local transition patterns, learned on the fly in a particular scene. Non-linear motion, historically observed in a coarse grid cell of the environment, informs the LSTM predictor.

Several authors used LSTMs to estimate kinodynamic motion of vehicles, combining the benefits of the physics-based and the pattern-based methods (Deo and Trivedi, 2018; Raipuria et al., 2018). Raipuria et al. (2018) augmented the LSTM model with the road infrastructure indicators, expressed in the curvilinear coordinate system, to better predict motion in curved road segments. Deo and Trivedi (2018) proposed an interaction-aware multiple-LSTM model to compute stochastic maneuver-dependent predictions of a vehicle, and augment it with an LSTM-based maneuver classification and mixing mechanism.

Other approaches used RNN as models of spatiotemporal graphs for problems that require both spatial and temporal reasoning (Dai et al., 2019; Eiffert and Sukkarieh, 2019; Huang et al., 2019; Ivanovic and Pavone, 2019; Jain et al., 2016; Vemula et al., 2018). Jain et al. (2016) proposed an approach for training sequence prediction models on arbitrary high-level spatiotemporal graphs, whose nodes and edges are represented by RNNs. The resulting graph is a feed-forward, fully differentiable, and jointly trainable RNN mixture. Vemula et al. (2018) applied this method to jointly predict transitions in human crowds.

RNN abilities for prediction of time-series is also combined with different neural networks architectures (Choi and Savarese, 2010; Li et al., 2019; Schmerling et al., 2018; Zhan et al., 2018; Zheng et al., 2016). Schmerling et al. (2018) considered a traffic weaving scenario and propose a conditional variational autoencoder (CVAE) with RNN subcomponents to model interactive human driver behaviors. The CVAE characterizes a multi-modal distribution over human actions at each time step conditioned on interaction history, as well as future robot action choices. Zheng et al. (2016) described a hierarchical policy approach that automatically reasons about both long-term and short-term goals. The model uses recurrent convolutional neural networks (CNNs) to make predictions for macro-goals (intermediate goals) and micro-actions (relative motion), which are trained independently by supervised learning, combined by an attention module, and finally jointly fine-tuned. Zhan et al. (2018) extended this approach using variational RNNs. Choi et al. (2019) used spatial–temporal graphs in combination with CVAE. The spatial–temporal graphs are used to model the relational influence among predicted agents. Conditions of the CVAE are represented by estimated intentions. In addition, Li et al. (2019) proposed a hierarchical architecture where an upper level (based on variational RNN) provides predictions of discrete coordination activities between agents and a lower level generates actual geometric predictions (using a conditional generative adversarial network (GAN)). The probabilistic framework called *multiple futures predictor* (MFP) (Tang and Salakhutdinov, 2019) models joint behavior of an arbitrary number of agents via a dynamic attention-based state encoder for capturing relationships between agents, a set of stochastic, discrete latent variables per agent to allow for multi-modal future behavior, as well as interactive and stepwise parallel rollouts with agent-specific RNNs to model future interactions. Furthermore, there model allows to make hypothetical rollouts under assumptions of behavior for a particular agent.

Several recent works (Jain et al., 2019; Radwan et al., 2018; Rhinehart et al., 2019; Ridel et al., 2019; Srikanth et al., 2019; van der Heiden et al., 2019; Xue et al., 2018; Zhao et al., 2019) combined the benefits of sequential (e.g., RNN-based) and convolutional approaches for modeling jointly the spatial and temporal relations of the observed agents' motion. Xue et al. (2018) introduced a hierarchical LSTM model, which combines inputs on three scales: trajectory of the person, social neighborhood and features of the global scene layout, extracted with a CNN. Zhao et al. (2019) proposed the multi-agent tensor fusion encoding, which fused contextual image of the environment with sequential trajectories of agents, thus retaining spatial relation between features of the environment and capturing interaction between the agents. This method is applied to both pedestrian and vehicles. In addition, Rhinehart et al. (2019) presented a prediction scheme for multi-agents that combines CNNs with a generative model based on RNNs. Moreover, the approach conditions the predictions on inferred intentions of the agents. Srikanth et al. (2019) proposed a novel input representation for learning vehicle dynamics, which includes semantics images, depth information, and other agents' positions. This input is projected into top-down view and fed into the autoregressive convolutional LSTM model to learn temporal dynamics. LSTMs have been also used to predict sequence of future human movements based on a learned reward map (Saleh et al., 2019).

Recently, many authors have applied the GAN architecture to achieve multi-modality in the prediction output

(Amirian et al., 2019; Gupta et al., 2018; Kosaraju et al., 2019). For instance, Gupta et al. (2018) extended the Social-LSTM by using GANs and a novel variety loss that encourages the generative network to produce diverse multi-modal predictions. Kosaraju et al. (2019) use graph attention network in combination with the GAN architecture to better capture relative importance of surrounding agents and semantic features of the environment.

4.2. Non-sequential models

Learning motion patterns in complex environments requires the model to generalize across non-uniform, context-dependent behaviors. Specifying causal constraints, e.g., through the Markovian assumption for the sequential models and additionally the particular functional form for the physics-based methods, might be too restrictive for these situations. Alternatively, instead of focusing on the local transitions of the system, *non-sequential approaches* aim to directly learn a distribution over long-term trajectories, that the observed agent may follow in the future, i.e., learn a set of full motion patterns from data.

Most basic non-sequential approaches are based on clustering the observed trajectories, which creates a set of long-term motion patterns (Bennewitz et al., 2005, 2002; Bera et al., 2016, 2017; Chen et al., 2008). IN this way the global structure of the workspace is imposed on top of a sequential model. Clustering-based approaches are illustrated in Figure 9(d). Bennewitz et al. (2005, 2002) cluster recorded trajectories of humans into global motion patterns using the EM algorithm and build an HMM model for each cluster. For prediction, the method compares the observed track with the learned motion patterns, and reasons about which patterns best explain it. Uncertainty is handled by probabilistic mixing of the most likely patterns. Similarly, Zhou et al. (2015) modeled the global motion patterns in a crowd with linear dynamic systems using EM for parameters estimation. Several authors (Makris and Ellis, 2002; Piciarelli et al., 2005) proposed graph structures to efficiently capture the branching of trajectory clusters. Chen et al. (2008) proposed a method for dynamic clustering of the observed trajectories, assuming that the set of complete motion patterns may not be available at the time of prediction, e.g., in new environments. Sung et al. (2012) proposed to represent the agent's states as short trajectories rather than static positions. This higher level of abstraction provides greater flexibility to represent not only position, but also velocity and intention. Suraj et al. (2018) directly used a large-scale database of observed trajectories (up to 10 million) to estimate the future positions of a vehicle given only its position, rotation, and velocity. Combining the concepts of local motion patterns and clustering, Carvalho et al. (2019) represent each cluster with a piece-wise linear vector field over an arbitrary state-space mesh.

Several approaches use GPs or mixture models as cluster centroids representation (Kim et al., 2011; Tay and Laugier, 2008; Yoo et al., 2016). Tay and Laugier (2008) introduced an approach to predict motion of a dynamic object in known scenes based on GMMs and GPs. Kim et al. (2011) modeled continuous dense flow fields from a sparse set of vector sequences. Yoo et al. (2016) proposed to learn most common patterns in the scene and their co-occurrence tendency using topic mixture and GMMs. Observed trajectories were clustered into several groups of typical patterns that occur at the same time with high probability. Given a set of observed trajectories, prediction was performed considering the dominant pattern group. Makansi et al. (2019) presented a mixture density network architecture, which generates multiple hypotheses of future position in fixed interval Δt and then fits a mixture of Gaussian or Laplace distributions to these hypothesis.

Clustering-based methods, discussed so far, generalize statistical information in a particular environment. In comparison, location-invariant methods, based on matching the observed partial trajectory to a set of prototypical trajectories, can be used in arbitrary free space (Hermes et al., 2009; Keller et al., 2011; Xiao et al., 2015), see Figure 9(e). Hermes et al. (2009) predicted trajectories of vehicles by comparing the observed track to a set of motion patterns, clustered with a rotationally invariant distance metric. In their probabilistic hierarchical trajectory matching (PHTM) approach, Keller et al. (2011) proposed a probabilistic search tree of sample human trajectory snippets to find the corresponding matching sub-sequence. Xiao et al. (2015) decomposed the set of sample trajectories into predefined motion classes, such as wandering or stopping, rotating and aligning them to start from the same point and have the longest span along the same axis. In contrast, skipping the clustering step, Nikhil and Tran Morris (2018) proposed a simple method to map the input trajectory of fixed length to the full future trajectory using a CNN.

For interaction-aware non-sequential motion prediction, several authors considered the case with two interacting agents (Käfer et al., 2010; Luber et al., 2012). Käfer et al. (2010) proposed a method for joint pairwise vehicle trajectory estimation at intersections. Comparing the observed motion pattern with those stored in a motion database, several prospective future trajectories were extracted independently for each vehicle. The probability of each pair of possible future trajectories was then estimated. Luber et al. (2012) modeled joint pairwise interactions between two people using social information. The authors learned a set of dynamic motion prototypes from observations of relative motion behavior of humans in public spaces. An unsupervised clustering technique determines the most likely future paths of two humans approaching a point of social interaction.

In contrast to multi-agent clustering, Trautman and Krause (2010) used GPs for making single-agent trajectory predictions. Then, an interaction potential re-weights the set of trajectories based on how close people are located to each other at every moment. A follow-up work (Trautman et al., 2013) incorporated goal information into the model: the goal position is added as a training point into the GP. Another approach by Su et al. (2017) used a social-aware LSTM-based crowd descriptor, which was later integrated into the deep GP to predict a complete distribution over future trajectories of all people.

Recently, several approaches for non-sequential prediction of vehicle motion using CNNs were presented (Cui et al., 2019; Djuric et al., 2018; Hong et al., 2019). An uncertainty-aware CNN-based vehicle motion prediction approach was presented by Djuric et al. (2018). The authors used a high-definition map image with projected prior motion of the target vehicle and full surrounding context as an input to the CNN, which produced the short-term trajectory of the target vehicle. The approach is extended by Cui et al. (2019) to inferring multi-modal predictions. Hong et al. (2019) proposed two methods for output representation using multi-modal regression with uncertainty or stacks of grid-map crops. Chai et al. (2019) used a fixed set of state-sequence "anchor" trajectories (clustered from training data), which correspond to possible modes of future behavior, as input to a CNN for mid-level scene features inference, and predict a discrete distribution over these anchors. For each anchor, the method regresses offsets from anchor waypoints along with uncertainties, yielding a Gaussian mixture at each time step.

5. Planning-based approaches

Planning-based approaches solve a sequential decision-making problem by reasoning about the future to infer a model of agent's motion. These approaches follow the *sense–reason–act* paradigm introduced earlier in Section 2. Unlike the previous two modeling approaches, the planning-based approach incorporates the concept of a rational agent when modeling human motions. By placing an assumption of rationality on the human, the models used to represent human motion must take into account the impact of current actions on the future as part of its model. As a result, much of the work covered in this section used objective functions that minimize some notion of the total cost of a sequence of actions (motions), and not just the cost of one action in isolation.

Here we classify planning-based approaches into two sub-categories, depicted in Figure 10. *Forward planning-based approaches* (Section 5.1) use a pre-defined cost function to predict human motion, and *inverse planning-based approaches* (Section 5.2) infer the cost (or policy) function from observations of human behavior and then use that cost (or policy) function to predict human motion.

5.1. Forward planning approaches

5.1.1. Motion and path planning methods. To make basic goal-informed predictions, several methods use optimal motion and path planning techniques with a hand-crafted cost-function (Bruce and Gordon, 2004; Gong et al., 2011; Vasishta et al., 2017; Xie et al., 2013; Yi et al., 2016). Bruce and Gordon (2004) proposed to use a path planning algorithm to infer how a person would move towards destinations in the environment. Predictions were performed using a set of learned goals. Gong et al. (2011) used multiple long-term goal-directed path hypotheses from different homotopy classes, generated with a modified A* algorithm (Bhattacharya et al., 2010). Xie et al. (2013) described a Dijkstra-based approach to predict human transitions across *dark energy* fields generated from video data. Every goal location generates an attractive *dark matter* Gaussian force field, whereas every non-walkable location generates a repulsive one. The dark matter functional objects, the map and the goals are inferred on-line using a Monte Carlo Markov chain technique. For predicting human motion in a crowd, Yi et al. (2016) introduced an energy map to model the traveling difficulty of each location in the scene, accounting for obstacles layout, moving people, and stationary groups. The energy map is personalized for each observed agent, and the fast marching method (FMM) (Sethian, 1996) was used to predict the person's path. Vasishta et al. (2017) used A* search over the potential cost-map function for pedestrian trajectory prediction, aiming to recognize illegal crossing intention of the observed agent. The potential field accounts for semantic properties of the urban environment.

Other methods model the probabilities of future motion based on cost-to-go value estimates (Best and Fitch, 2015; Karasev et al., 2016; Rudenko et al., 2017; Vasquez, 2016; Yen et al., 2008). Yen et al. (2008) proposed a probabilistic goal-directed motion model that accounts for several goals in the environment. The method computes the cost-to-go function for each goal and evaluates the probabilities of feasible transitions in each state. A person's trajectory is predicted using a particle filter with Monte Carlo sampling. Best and Fitch (2015) proposed a Bayesian framework that exploits the set of path hypotheses to estimate the intended destination and the future trajectory. To this end, a probabilistic dynamical model was used, which evaluates the next states of the agent based on the decrease of the distance to the intended goal. Hypotheses are generated from the probabilistic roadmap (PRM). Karasev et al. (2016) solve the prediction problem using a jump-Markov decision process (MDP), modeling the agents' behavior as switching non-linear dynamical systems. A soft MDP policy describes the nonlinear motion dynamics, and the latent goal variable governs the switches. The method uses hand-crafted costs for each surface type (e.g., sidewalk, crosswalk, road, grass), and handles time-dependent information such as traffic signals. Instead of using an MDP formulation, Vasquez (2016) proposed the FMM to compute the cost-to-go function for a set of goals. The predictor used a velocity-dependent probabilistic motion model, described the temporal evolution along the predicted path,

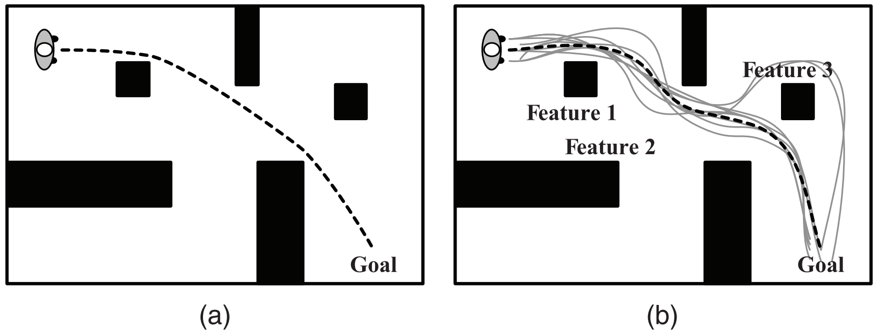

Fig. 10. Examples of the planning-based approaches: (a) forward planning approach, which used a predefined cost function (e.g., Euclidean distance), and (b) inverse planning approach, which infers the feature-based cost function from observations.

and offered a gradient-based goal prediction that allows quick recognition of the intended destination changes.

5.1.2. Multi-agent forward planning. Most planning-based methods discussed so far do not consider interactions between agents in the scene. To account for the presence of other agents, several authors proposed to modify individual optimal policies locally with physics-based methods (Rudenko et al., 2018a; van Den Berg et al., 2008; Wu et al., 2018) or imitation learning Muench and Gavrila (2019). A crowd simulation approach that combines global planning and local collision avoidance was presented by van Den Berg et al. (2008). A global path for each agent is computed using a PRM, considering only static obstacles. Local collision avoidance along the global path is done jointly for all agents using the RVO (van den Berg et al., 2008) method. Rudenko et al. (2018a) extended the MDP-based approaches (Karasev et al., 2016; Ziebart et al., 2009) with a fast random-walk-based method to generate joint predictions for all observed people using social forces. The authors extended their approach considering group-based social motion constraints in Rudenko et al. (2018b). Wu et al. (2018) extended the gridmap transition-based and reachability-based framework (Coscia et al., 2018; Rehder and Klöden, 2015) with automatic inference of local goal points, and calculate the stochastic policy in each cell, augmenting the physics-based dynamics with optimal motion direction. The motion of pedestrians is predicted jointly with other traffic participants by risk checking of future states based on gap acceptance model (Brewer et al., 2006). Instead of using a physics-based approach (e.g., social forces) for augmenting the MDP-based predictor, Muench and Gavrila (2019) proposed to learn an additional interaction-aware Q-function with imitation learning.

A number of approaches considered cooperative planning in joint state-space that includes all agents (Bahram et al., 2016; Broadhurst et al., 2005; Chen et al., 2017; Rösmann et al., 2015). Broadhurst et al. (2005) used Monte Carlo sampling to generate probability distributions over future trajectories of the vehicles and pedestrians jointly. The approach considers several available actions for each agent in the scene: each vehicle executes one of the hand-crafted behaviors, and humans are assumed to move freely in all directions. In addition, Rösmann et al. (2017) considered planning for cooperating agents. A set of topologically distinct candidate trajectories for each person is computed using trajectory optimization techniques (Rösmann et al., 2015). Among those trajectories, the best candidate is chosen according to a metric that includes group integrity, right versus left motion bias, and curvature constraints. Finally, the encounter is resolved jointly in an iterative fashion. The interaction point of minimal spatial separation is computed between each two people, who adjust their trajectories accordingly, possibly switching to a different topological candidate. Mavrogiannis and Knepper (2016) represented multi-agent interaction through the use of braid groups (topological patterns) that formalize trajectories sets. At inference time, the problem of predicting joint trajectories is posed as a graph search in a permutation graph.

Joint planning for the robot and the human is addressed by several works (Bandyopadhyay et al., 2013; Chen et al., 2017; Galceran et al., 2015). Assuming availability of a fixed set of goals, Bandyopadhyay et al. (2013) solved an optimal motion problem for each of it, and generated appropriate motion policies. The latter were used to estimate the future evolution of the joint state-space of the robot and the human. Galceran et al. (2015) introduced a multi-policy decision-making system to generate robot motions based on predicted movements of other agents in the scene, estimated with a changepoint-based technique (Fearnhead and Liu, 2007). Likelihood of future actions were sampled from the policies. The final prediction was generated by an exhaustive search of closed-loop forward simulations of these samples. The approach is well suited for predicting future macro-actions (i.e., turn left or right, slow down or speed up). Bahram et al. (2016) generated joint robot and agents' motions using a sequential game theory technique. The approach presented an interactive prediction and planning loop where a sequence of predictions (i.e., motion primitives) was generated for the ego-vehicle by considering the sequential evolution of the entire scene. Chen et al. (2017) developed a de-centralized multi-agent collision avoidance algorithm, which resolves local interactions with a learned joint value function that implicitly encodes cooperative behaviors.

5.2. Inverse planning approaches

Forward planning approaches, discussed so far, make an explicit assumption about the optimality criteria (reward or cost function) of an agent's motion. In this section, we discuss algorithms that estimate the reward function of agents (or directly a policy) from observations, using statistical and imitation learning techniques (for a survey on imitation learning techniques applied to robotic systems, we refer the reader to Osa et al. (2018)). Inverse planning methods assume that the reward or cost function, which depends on contextual and social features and defines the rational behavior, can be learned from observations (see Figure 10(b)).

5.2.1. Single-agent inverse planning. In their influential work, Ziebart et al. (2009) proposed to learn a reward function yielding goal-directed behavior of pedestrians using maximum entropy (MaxEnt) inverse optimal control (IOC). Humans are assumed to be near-optimal decision makers with stochastic policies, learned from observations, which are used to predict motion as a probability distribution over trajectories. Building upon the work of Ziebart et al. (2009), Kitani et al. (2012) expanded it to include the labeled semantic map of the environment. An IOC method takes the semantic map as an input, and learns the feature-based cost function that captures agents' preferences for, e.g., walking on the sidewalk, or keeping some distance from parked cars. Previtali et al. (2016) proposed an approach that adopts linear programming formulation of IRL. Using a discrete and non-uniform representation of the 2D workspace, it scales linearly with respect to the size of the environment. Chung and Huang (2010) presented an MDP-based model that describes spatial effects between agents and the environment. The authors used IRL to estimate cost of each state as a linear combination of trajectory length, static and dynamic obstacle avoidance and steering smoothness. Special context-based spatial effects (SSEs) are identified by comparing the costs of the states, learned with IRL, and the actual observed trajectories. A follow-up work (Chung and Huang, 2012) introduced a feature-based representation of SSEs, which can be modeled before being naturally observed, as in Chung and Huang (2010).

Instead of IRL, other works used different techniques to learn the reward function (Huang et al., 2016; Rehder et al., 2018). Rehder et al. (2018) solved the problem of intention recognition and trajectory prediction in one single ANN. The destinations and costly areas are predicted from stereo images using a recurrent mixture density network (RMDN). Planning towards these destinations is performed using fully CNNs. Two different architectures for planning are proposed: an MDP network and a forward–backward network, both using contextual features of the environment. Huang et al. (2016) proposed an approach that exploits two CNNs to learn a reward function considering spatial and temporal contextual information from a video sequence. A SMN learns the spatial context of human motion. An orientation network (ON) is used to model the position variation of the object. The Dijkstra algorithm is used to find the minimum cost solution over a graph whose edges' weights are set by considering the reward function and the facing orientation computed by the two networks (SMN and ON).

All the detailed methods show that IRL or similar methods are providing powerful tools to learn human behaviors. Furthermore, Shen et al. (2018) showed that under some particular requirements (i.e., when the feature vector, model parameter, and output representation are invariant under a rigid-body transformation of the world fixed coordinate frame), IRL is suitable for learning location-independent transferable motion models.

5.2.2. Imitation learning. Instead of first learning a reward function and then applying planning techniques to generate motion predictions, imitation learning approaches directly extract a policy from the data. The GAIL approach, proposed by Ho and Ermon (2016), aims for matching long-term distributions over states and actions. It uses a GAN-based (Goodfellow et al., 2014) optimization procedure, in which a discriminator tries to distinguish between observations from experts and generated ones by making model rollouts. Afterwards, a model is trained to make predictions that yield similar long-term distributions over states and actions. This method has been successfully applied to learning human highway driving behavior (Kuefler et al., 2017) and training joint pedestrian motion models (Gupta et al., 2018). Li et al. (2017) extended GAIL by introducing a component to the loss function, which maximizes the mutual information between the latent structure and observed trajectories. They test their approach in a simulated highway driving scenario, predicting the driver's actions given an input image and auxiliary information (e.g., velocity, last actions, damage), and show that it is able to imitate human driving, while automatically distinguishing between different types of behaviors.

Differently from GAIL, the deep generative technique by Rhinehart et al. (2018a) adopts a fully differentiable model, which is easy to train without the need of an expensive policy gradient search. By minimizing a symmetrized cross-entropy between the distributions of the policy and of the demonstration data, the method allows to learn a policy that generates predictions which balance precision (i.e., avoid obstacle areas) and diversity (i.e., being multi-modal).

5.2.3. Multi-agent inverse planning. In the following, we review several inverse planning approaches that predict multi-agent motions (Fernando et al., 2019; Kretzschmar et al., 2014; Kuderer et al., 2012; Lee et al., 2017; Ma et al., 2017; Pfeiffer et al., 2016). Kuderer et al. (2012) and Kretzschmar et al. (2014) proposed a continuous formulation of the MaxEnt IOC (Ziebart et al., 2009) by considering a continuous spline-based trajectory representation. Their method relies on several features (e.g., travel time, collision avoidance) to capture physical and topological aspects of the pedestrians trajectories. Pfeiffer et al. (2016) extended the latter works by introducing the variable end position of each trajectory, thus reasoning over the agents' goals. Walker et al. (2014) presented an unsupervised learning approach for visual scene prediction. The approach exploits mid-level elements (i.e., image patches) as building blocks for jointly predicting positions of agents in the scene and changes in their visual appearance. The learned reward function defines the probability of a patch moving to a different location in the image. To generate predictions, the method performs a Dijkstra search on the learned reward function considering several goals. Ma et al. (2017) combine the fictitious play (Brown, 1951) game theory method with the deep learning-based visual scene analysis. Future paths hypothesis are generated jointly and iteratively: each pedestrian adapts their motion based on the predictions of the other pedestrians' actions. IRL's reward function features encode social compliance, neighborhood occupancy, distance to the goal, and body orientation. Gender and age attributes, extracted with a deep network from video, define the possible average velocity of pedestrians.

Lee et al. (2017) formulated the prediction problem as an optimization task. The method reasons on multi-modal future trajectories accounting for agent interactions, scene semantics, and expected reward function, learned using a sampling-based IRL scheme. The model is wrapped into the single end-to-end trainable RNN encoder-decoder network called DESIRE. The RNN architecture allows incorporation of past trajectory into the inference process, which improves prediction accuracy compared with the standard IRL-based techniques.

The previously discussed approaches for joint prediction assume multi-agent settings with rational and cooperative behavior of all agents. Differently, several approaches (Henry et al., 2010; Lee and Kitani, 2016) address the problem by modeling one target person as a rational agent, acting in a dynamic environment. The influence of other agents then becomes part of the stochastic transition model of the environment. For example, Henry et al. (2010) proposed an IRL-based method for imitating human navigation in crowded environments. They conjectured that humans take into account the density and velocity of nearby people and learned a reward function that weights between these and additional features. Another approach by Lee and Kitani (2016) learned a reward function that explains the behavior of a wide receiver in American football, whose strategy takes into account the behavior of the defenders. Models of the dynamic environment (e.g., linear or GPs) are used as transitions in the IRL framework.

Rhinehart et al. (2019) has developed a multi-agent forecasting model called estimating social-forecast probabilities (ESP) that uses exact likelihood inference (unlike VAEs or GANs) derived from a deep neural network for forecast trajectories. In contrast to most standard trajectory forecasting methods, the approach is able to reason conditionally based on additional information that it was not trained to use by accepting agent goals at test time. The approach uses a generative multi-agent model in order to perform prediction conditioned on goals (PRECOG).

6. Contextual cues

In this section, we discuss the categorization of the contextual cues, in those dealing with the target agent (Section 6.1), the other dynamic agents (Section 6.2), and the static environment (Section 6.3).

6.1. Cues of the target agent

Most essential cues, used to predict future states of an agent, are related to the agent itself. To this end most of the algorithms use current position and velocity of the target agent (Bahram et al., 2016; Bansal et al., 2019; Bennewitz et al., 2005; Bera et al., 2016; Elfring et al., 2014; Ferrer and Sanfeliu, 2014; Habibi et al., 2018; Karasev et al., 2016; Kitani et al., 2012; Kucner et al., 2017; Kuderer et al., 2012; Luo and Cai, 2019; Pellegrini et al., 2009; Rhinehart et al., 2018a; Rudenko et al., 2018b; Trautman and Krause, 2010; Wu et al., 2018; Ziebart et al., 2009), often considering also the history of recent states/velocities. Position and velocity are also the main attributes of the target agent in vehicle motion prediction tasks (Broadhurst et al., 2005; Hermes et al., 2009; Käfer et al., 2010). Considering the head orientation or full articulated pose of the person (Blaiotta, 2019; Hasan et al., 2018; Kooij et al., 2019, 2014; Mínguez et al., 2018; Quintero et al., 2014; Roth et al., 2016; Schulz and Stiefelhagen, 2015; Unhelkar et al., 2015) may bring valuable insights on the target agent's immediate intentions or their awareness of the environment. Considering additional semantic attributes of the target agent may further refine the quality of predictions: gender and age in Ma et al. (2017), personality type (Bera et al., 2017), class of the dynamic agent (e.g. a person or a cyclist in pedestrian areas, motorcycle, car, or a truck on a highway) (Altché and de La Fortelle, 2017; Ballan et al., 2016; Coscia et al., 2018), person's attention and awareness of the robot's presence in Oli et al. (2013), Kooij et al. (2019), and Blaiotta (2019), and raised arm as a bending intention indicator for cyclists (Kooij et al., 2019; Pool et al., 2019).

6.2. Cues of other dynamic agents

Most of the time all agents navigate in a shared environment, adapting their actions, timing, and route based on the others' presence and behavior. Therefore, for predicting motion it is beneficial to consider interaction between moving agents. We classify the existing approaches in three categories: *unaware predictors*, *individual-aware predictors*, and *group-aware predictors*.

The class of unaware predictors includes all methods that generate motion prediction for a single agent, considering only the static contextual cues of the environment. Having no need to explicitly define or learn the interaction model, these methods are simpler to set up, require less training data to generalize, typically have less parameters to estimate. Simpler physics-based methods, such as linear velocity projection or constant acceleration models, are unaware predictors (Bai et al., 2015; Coscia et al., 2018; Elnagar, 2001; Elnagar and Gupta, 1998; Foka and Trahanias, 2010; Koschi et al., 2018; Vasishta et al., 2017, 2018; Xie et al., 2018; Zhu, 1991). Many pattern-based (Bennewitz et al., 2005, 2002; Carvalho et al., 2019; Chen et al., 2016, 2008; Goldhammer et al., 2014; Habibi et al., 2018; Han et al., 2019; Hermes et al., 2009; Huynh and Alaghband, 2019; Kim et al., 2017, 2011; Kucner et al., 2017, 2013; Makansi et al., 2019; Makris and Ellis, 2002; Molina et al., 2018; Nikhil and Tran Morris, 2018; Piciarelli et al., 2005; Ridel et al., 2019; Saleh et al., 2018b; Sung et al., 2012; Suraj et al., 2018; Tadokoro et al., 1993; Thompson et al., 2009; Unhelkar et al., 2015; Wang et al., 2016; Xiao et al., 2015; Xue et al., 2019, 2017) and planning-based methods (Gong et al., 2011; Karasev et al., 2016; Kitani et al., 2012; Rhinehart et al., 2018a; Rudenko et al., 2017; Vasquez, 2016; Yen et al., 2008; Ziebart et al., 2009) are unaware predictors, owing to the increase of complexity for conditioning the learned transition patterns or optimal actions on the presence and positions of other agents. Methods for predicting pedestrians crossing behavior (Gu et al., 2016; Keller and Gavrila, 2014; Kooij et al., 2014; Mínguez et al., 2018; Quintero et al., 2014; Roth et al., 2016; Schulz and Stiefelhagen, 2015) and cyclist motion (Pool et al., 2019, 2017; Saleh et al., 2018a; Zernetsch et al., 2016) typically treat each agent individually.

Individual-aware predictors methods consider the interaction between agents by modeling or learning their influence on each other. Physics-based methods that use social forces (Blaiotta, 2019; Elfring et al., 2014; Ferrer and Sanfeliu, 2014; Karamouzas et al., 2009; Luber et al., 2010; Oli et al., 2013; Zanlungo et al., 2011) or similar local interaction models (Bansal et al., 2019; Karamouzas and Overmars, 2010; Kim et al., 2011; Luo and Cai, 2019; Paris et al., 2007; Pellegrini et al., 2009, 2010; Pettré et al., 2009; Robicquet et al., 2016; Yamaguchi et al., 2011) are classical examples of individual-aware prediction models. A pattern-based approach by Ikeda et al. (2012) models deviations from the desired path using social forces. In general, however, learning joint motion patterns is a considerably harder task. For example, the approach of Trautman and Krause (2010); Trautman et al. (2013) learned unaware motion patterns, and then evaluated the predicted probability distribution over the joint paths using an explicit interaction potential. Luber et al. (2012) learned pairwise joint motion patterns of two humans approaching the spatial point of interaction. The approach by Yoo et al. (2016) learns which motion patterns are likely to occur at the same time and uses this information for predicting the future motion of several dynamic objects. Some approaches propose to learn a motion policy or reward function that accounts for dynamic objects in the surrounding (Chung and Huang, 2010, 2012; Henry et al., 2010; Lee and Kitani, 2016; Vemula et al., 2017). Rudenko et al. (2018a) proposed an MDP planning-based method, where optimal policies of people are locally modified to account for other dynamic entities. Wu et al. (2018) and Zechel et al. (2019) discounted predicted transition probabilities to states in collision with other agents. Muench and Gavrila (2019) decomposed the interactive planning problem into two policies with the corresponding Q-functions: one for prediction in static environment, and another for interaction prediction in an obstacle-free environment. Many deep learning methods consider interactions between participants: explicitly modeling interacting entities (Alahi et al., 2016; Amirian et al., 2019; Bartoli et al., 2018; Choi et al., 2019; Eiffert and Sukkarieh, 2019; Fernando et al., 2018, 2019; Gupta et al., 2018; Hasan et al., 2018; Huang et al., 2019; Ivanovic and Pavone, 2019; Kosaraju et al., 2019; Pei et al., 2019; Pfeiffer et al., 2018; Radwan et al., 2018; Rhinehart et al., 2019; Sadeghian et al., 2019; Saleh et al., 2019; Shi et al., 2019; Su et al., 2017; van der Heiden et al., 2019; Varshneya and Srinivasaraghavan, 2017; Vemula et al., 2018; Xu et al., 2018; Xue et al., 2018; Zhao et al., 2019), implicitly as a result of pixel-wise prediction (Walker et al., 2014), or by learning a joint motion policy (Lee et al., 2017; Ma et al., 2017; Shalev-Shwartz et al., 2016; Zhan et al., 2018). Many vehicle prediction methods consider interaction between traffic participants (e.g., Agamennoni et al., 2012; Altché and de La Fortelle, 2017; Bahram et al., 2016; Broadhurst et al., 2005; Chai et al., 2019; Cui et al., 2019; Dai et al., 2019; Deo and Trivedi, 2018; Ding et al., 2019; Djuric et al., 2018; Hong et al., 2019; Jain et al., 2019; Käfer et al., 2010; Kim et al., 2017; Kuhnt et al., 2016; Li et al., 2019; Park et al., 2018; Raipuria et al., 2018; Srikanth et al., 2019). Kooij et al. (2019) considered whether the ego-vehicle is on a potential collision course when predicting the road user path in their SLDS-based approach.

Group-aware predictors also recognize affiliations and relations of individual agents and respect the probability of them traveling together, as well as model an appropriate reaction of other agents to the moving group formation. For example, several physics-based methods model group relations by introducing additional attractive forces between group members (Choi and Savarese, 2010; Karamouzas and Overmars, 2012; Moussaïd et al., 2010; Pellegrini et al., 2010; Qiu and Hu, 2010; Robicquet et al., 2016; Seitz et al., 2012; Singh et al., 2009; Yamaguchi et al., 2011). Several learning-based approaches that use LSTMs (Alahi et al., 2016; Bartoli et al., 2018; Pfeiffer et al., 2018; Shi et al., 2019; Varshneya and Srinivasaraghavan, 2017; Zhang et al., 2019) may be capable of implicitly learning intra- and inter-group coherence behavior, however only the work by Bisagno et al. (2018) states this capability explicitly. A planning-based approach that implicitly respects group integrity by increasing the costs of passing between group members was presented by Rösmann et al. (2017) and an approach that explicitly models group motion constraints was presented by Rudenko et al. (2018b).

Algorithms using high-level context information about dynamic agents produce more precise predictions in a variety of cases. Learning advanced social features of human motion improves interactive predictors performance, for instance different parameters for interactions of heterogeneous agents (Ferrer and Sanfeliu, 2014), advanced motion criteria such as *social comfort* of navigation (Kuderer et al., 2012; Luber et al., 2012; Pfeiffer et al., 2016) or “desire to move with the flow” or “avoid dense areas” (Henry et al., 2010). Some approaches model prior knowledge in terms of the dynamics of moving agents (Lee et al., 2017; Rösmann et al., 2017), human attributes and personal traits (Ma et al., 2017). Chung and Huang (2012) presented a general framework for learning context-related spatial effects, which affect the human motion, such as avoiding going through a waiting line, or in front of a person, who observes the work of art in a museum.

Modeling also the influence of the robot’s presence on the agents’ paths is another interesting line of research: Trautman and Krause (2010) and Oli et al. (2013) tackled this problem by placing the robot as a peer-interacting agent among moving humans. Several authors (Kretzschmar et al., 2014; Kuderer et al., 2012; Pfeiffer et al., 2016; Rösmann et al., 2017) optimized joint trajectories for all humans and the robot. A relevant case of modeling the effect of robotic herd actions on the location and shape of the flock of animals was studied by Sumpter and Bulpitt (2000). Similarly, Schmerling et al. (2018) condition human response on the candidate robot actions for modeling pairwise human–robot interaction. Eiffert and Sukkarieh (2019) included the robot as an interacting agent in the LSTM-based predictor. Tang and Salakhutdinov (2019) computed a conditional probability density over the trajectories of other agents given the hypothetical rollout for the robot.

6.3. *Cues of the static environment*

Humans adapt their behaviors according not only to the movements of the other agents but also to the environment’s shape and structure, making extensive use of its topology to reason on the possible paths to reach the long-term goal. Many existing prediction algorithms make use of such geometric information of the environment.

Some approaches produce *unaware predictions*, assuming an obstacle-free environment. This category includes several physics-based approaches (Bai et al., 2015; Blaiotta, 2019; Elnagar, 2001; Elnagar and Gupta, 1998; Foka and Trahanias, 2010; Pettré et al., 2009; Schneider and Gavrila, 2013; Zhu, 1991). Pattern-based methods usually model obstacles implicitly, by learning collision-free patterns (Bennewitz et al., 2002; Carvalho et al., 2019; Chen et al., 2016, 2008; Ellis et al., 2009; Ferguson et al., 2015; Han et al., 2019; Hasan et al., 2018; Huynh and Alaghband, 2019; Jacobs et al., 2017; Joseph et al., 2011; Kim et al., 2011; Kruse and Wahl, 1998; Kucner et al., 2017, 2013; Makansi et al., 2019; Makris and Ellis, 2002; Molina et al., 2018; Piciarelli et al., 2005; Saleh et al., 2018a,b; Sun et al., 2018; Sung et al., 2012; Tadokoro et al., 1993; Tay and Laugier, 2008; Thompson et al., 2009; Vasquez et al., 2008; Wang et al., 2015, 2016; Xue et al., 2019, 2017; Yoo et al., 2016). When facing a change in the obstacles’ configuration, such patterns become obstacle-unaware. Location-independent motion patterns are usually obstacle-unaware (Goldhammer et al., 2014; Hermes et al., 2009; Luber et al., 2012; Nikhil and Tran Morris, 2018; Unhelkar et al., 2015; Xiao et al., 2015). Pedestrian crossing prediction methods typically assume obstacle-free environment (Gu et al., 2016; Keller and Gavrila, 2014; Kooij et al., 2019, 2014; Mínguez et al., 2018; Quintero et al., 2014; Roth et al., 2016; Schulz and Stiefelhagen, 2015), as well as most of the vehicle prediction methods (Altché and de La Fortelle, 2017; Deo and Trivedi, 2018; Ding et al., 2019; Kim et al., 2017; Park et al., 2018; Raipuria et al., 2018; Suraj et al., 2018), which assume the road-surface to be free of static obstacles. Finally, many methods consider only dynamic entities, but no static obstacles in the environment (Alahi et al., 2016; Althoff et al., 2013, 2008b; Amirian et al., 2019; Bahram et al., 2016; Bartoli et al., 2018; Bera et al., 2016; Bisagno et al., 2018; Broadhurst et al., 2005; Dai et al., 2019; Eiffert and Sukkarieh, 2019; Fernando et al., 2018, 2019; Gupta et al., 2018; Huang et al., 2019; Ivanovic and Pavone, 2019; Käfer et al., 2010; Kim et al., 2015; Kuderer et al., 2012; Li et al., 2019; Pei et al., 2019; Pfeiffer et al., 2018; Radwan et al., 2018; Shi et al., 2019; Su et al., 2017; Trautman and Krause, 2010; Trautman et al., 2013; Varshneya and Srinivasaraghavan, 2017; Vemula et al., 2017; Vemula et al., 2018; Xu et al., 2018; Zanlungo et al., 2011; Zhang et al., 2019).

In several approaches the exact pose of the objects is known and utilized to compute more informed predictions (we refer to such methods as *obstacle-aware* methods). Mainly the social force-based and similar techniques model the interaction between the moving agents and individual static obstacles (Elfring et al., 2014; Ferrer and Sanfeliu, 2014; Karamouzas et al., 2009; Karamouzas and Overmars, 2010; Karasev et al., 2016; Kretzschmar et al., 2014; Luber et al., 2010; Luo and Cai, 2019; Oli et al., 2013; Paris et al., 2007; Pellegrini et al., 2009, 2010; Rhinehart et al., 2019; Robicquet et al., 2016; van den Berg et al., 2008; Yamaguchi et al., 2011; Zechel et al., 2019). Several location-independent pattern-based methods (Antonini et al., 2006; Aoude et al., 2011) can handle static object avoidance.

Still, obstacle-aware methods may fail in very cluttered environments, owing to the complexity of representing an environment with a set of individual obstacles. To overcome this difficulty, many prediction approaches use maps that are a more complete representation of the environment (we call them *map-aware* methods). Occupancy grid maps are the most common representation for these approaches, e.g., in the physics-based approach by Rehder and Klöden (2015) reachability-based transitions are calculated on a binary grid-map. In particular, the planning-based approaches use this kind of representation: thanks to the map they can infer global, intentional behaviors of the agents (Best and Fitch, 2015; Bruce and Gordon, 2004; Chen et al., 2017; Chung and Huang, 2010, 2012; Gong et al., 2011; Henry et al., 2010; Ikeda et al., 2012; Liao et al., 2003; Pfeiffer et al., 2016; Previtali et al., 2016; Rösmann et al., 2017; Rudenko et al., 2017, 2018a,b; Vasquez, 2016; Xie et al., 2013; Yen et al., 2008; Yi et al., 2016; Ziebart et al., 2009). Figure 7 shows the difference between the *pure motion-based predictions*, the *obstacle-aware*, and the *map-aware* approaches. The latter perform better in terms of global obstacle avoidance behavior during prediction.

Semantic map-based approaches extend the map-aware approaches by considering various semantic attributes of the static environment. A semantic map (Ballan et al., 2016; Coscia et al., 2018; Karasev et al., 2016; Kitani et al., 2012; Muench and Gavrila, 2019; Rehder et al., 2018; Rhinehart et al., 2018a; Rhinehart et al., 2018b; Ridel et al., 2019; Saleh et al., 2019; Shen et al., 2018; Tadokoro et al., 1993; van der Heiden et al., 2019; Vasishta et al., 2017, 2018; Zhao et al., 2019) or extracted features from a top-down image (Kosaraju et al., 2019; Sadeghian et al., 2019; Tang and Salakhutdinov, 2019; Xue et al., 2018) can be used to capture people preferences in walking on a particular type of surfaces. Furthermore, planning-based methods often use prior knowledge on potential goals in the environment (Best and Fitch, 2015; Karasev et al., 2016; Previtali et al., 2016; Rudenko et al., 2017; Vasquez, 2016). Location- and time-specific information in the particular environment may help to improve prediction quality (Molina et al., 2018; Sun et al., 2018).

Owing to the high level of structure in the environment, methods in autonomous driving scenarios extensively use available semantic information, such as street layout and traffic rules (Agamennoni et al., 2012; Chai et al., 2019; Choi and Savarese, 2010; Cui et al., 2019; Djuric et al., 2018; Gu et al., 2016; Hong et al., 2019; Jain et al., 2019; Keller and Gavrila, 2014; Kooij et al., 2014; Kuhnt et al., 2016; Lee et al., 2017; Petrich et al., 2013; Pool et al., 2019, 2017; Srikanth et al., 2019; Xie et al., 2018) or current state of the traffic lights (Gu et al., 2016; Jain et al., 2019; Karasev et al., 2016), also for predicting pedestrian and cyclist motion (Habibi et al., 2018; Kooij et al., 2019; Koschi et al., 2018).

7. Motion prediction evaluation

An important challenge for motion prediction methods is the design of experiments to evaluate their performance with respect to other methods and the requirements from the targeted application. In this section, we review and discuss common metrics and datasets to this end.

7.1. Performance metrics

Owing to the stochastic nature of human decision making and behavior, exact prediction of trajectories is rarely possible, and we require measures to quantify the similarity between predicted and actual motion. Different prediction types (see Figure 2) require different measures: for single trajectories we need geometric measures of trajectory similarity or final displacement, for parametric and non-parametric distributions over trajectories we can use geometric measures as well as difference measures for probability distributions. Metrics, commonly used in the literature, are summarized in Table 1.

7.1.1. Geometric accuracy metrics. Geometric measures are the most commonly used across all application domains. Several surveys have considered the topic of trajectory analysis and comparison (Morris and Trivedi, 2008; Pan et al., 2016; Quehl et al., 2017; Zhang et al., 2006; Zheng, 2015) where, based on the previous ones, only the recent survey by Quehl et al. (2017) specifically considers geometric similarity measures for trajectory prediction evaluation. In addition to that, we review the probabilistic metrics and the assessment of distributions with geometric methods in Section 7.1.2, and the experiments to evaluate robustness in Section 7.1.3.

Summarizing Morris and Trivedi (2008) and Quehl et al. (2017), we consider eight metrics as follows.

Mean Euclidean distance (MED), also called the *average displacement error* (ADE), averages Euclidean distances between points of the predicted trajectory and the ground truth that have the same temporal distance from their respective start points. An alternate form computes MED in a subspace between coefficients of the trajectories' principal components (PCA-Euclid). A third variant (MEDP) is a path measure able to compare paths of different length. For each (x, y)-point of the predicted path, the nearest ground-truth point is searched. Being a path measure, MEDP is invariant to velocity differences and temporal misalignment but does not account for temporal ordering. A fourth variant (n-ADE) measures MED only on non-linear segments of trajectories. MED measures are widely used by many authors across all domains, see Table 1. Many authors evaluate probabilistic predictions by computing expected MED under the predictive distribution, referring to it as *mean ADE*, *weighted mean ADE*, or,

Table 1. Metrics to evaluate motion prediction.

	Metric	Used by
Geometric	Average displacement error (ADE)	Pellegrini et al. (2009), Yamaguchi et al. (2011), Alahi et al. (2016), Sun et al. (2018), Bartoli et al. (2018), Vemula et al. (2017), Karasev et al. (2016), Kim et al. (2015), Vasquez et al. (2008), Yi et al. (2016), Rösmann et al. (2017), Yoo et al. (2016), Schulz and Stiefelhagen (2015), Zernetsch et al. (2016), Pool et al. (2017), Mínguez et al. (2018), Wu et al. (2018), Hermes et al. (2009), Raipuria et al. (2018), Deo and Trivedi (2018), Kim et al. (2017), Vemula et al. (2018), Radwan et al. (2018), Pfeiffer et al. (2018), Kooij et al. (2019), Quintero et al. (2014), Saleh et al. (2018a,b), Bisagno et al. (2018), Xue et al. (2019), Zhang et al. (2019), Shi et al. (2019), Zhao et al. (2019), Xue et al. (2017), Hasan et al. (2018), Xue et al. (2018), Su et al. (2017), Srikanth et al. (2019), Sadeghian et al. (2019), Park et al. (2018), Djuric et al. (2018), Xie et al. (2018), Gupta et al. (2018), Huynh and Alaghband (2019), Nikhil and Tran Morris (2018), Xu et al. (2018), Fernando et al. (2018), Cui et al. (2019), Luo and Cai (2019), Hong et al. (2019), Pei et al. (2019), Altché and de La Fortelle (2017), Huang et al. (2019), Chai et al. (2019), Amirian et al. (2019), Blaiotta (2019), Dai et al. (2019), Kosaraju et al. (2019), Ivanovic and Pavone (2019), Eiffert and Sukkarieh (2019), Saleh et al. (2019), Choi et al. (2019), Rhinehart et al. (2018a), Fernando et al. (2019), Li et al. (2019), Jain et al. (2019)
	Final displacement error (FDE)	Varshneya and Srinivasaraghavan (2017), Alahi et al. (2016), Vemula et al. (2017), Chung and Huang (2010), Vemula et al. (2018), Radwan et al. (2018), Bisagno et al. (2018), Xue et al. (2019), Zhang et al. (2019), Shi et al. (2019), Zhao et al. (2019), Xue et al. (2017), Hasan et al. (2018), Xue et al. (2018), Su et al. (2017), Sadeghian et al. (2019), Gupta et al. (2018), Huynh and Alaghband (2019), Nikhil and Tran Morris (2018), Xu et al. (2018), Fernando et al. (2018), Luo and Cai (2019), Pei et al. (2019), Huang et al. (2019), Amirian et al. (2019), Blaiotta (2019), Kosaraju et al. (2019), Ivanovic and Pavone (2019), Eiffert and Sukkarieh (2019), Choi et al. (2019)
	Modified Hausdorff distance (MHD)	Vasquez (2016), Kitani et al. (2012), Jacobs et al. (2017), Rudenko et al. (2017, 2018a,b), Yoo et al. (2016), Coscia et al. (2018), Shen et al. (2018), Habibi et al. (2018), Fernando et al. (2019), Saleh et al. (2019)
	Prediction accuracy (PA)	Ferrer and Sanfeliu (2014), Ikeda et al. (2012), Bera et al. (2016), Best and Fitch (2015), Ding et al. (2019), Hong et al. (2019)
Probabilistic	Negative log likelihood	Coscia et al. (2018), Rudenko et al. (2017), Suraj et al. (2018), Jain et al. (2019), Chai et al. (2019), Pool et al. (2019), Makansi et al. (2019), Ivanovic and Pavone (2019), Rhinehart et al. (2019)
	Negative log loss	Ma et al. (2017), Previtali et al. (2016), Vasquez (2016), Kitani et al. (2012), Tang and Salakhutdinov (2019)
	Predicted probability (PP)	Kooij et al. (2019, 2014), Rehder and Klöden (2015), Rudenko et al. (2018a,b)
	Minimum average or final displacement error (mADE, mFDE)	Lee et al. (2017), Park et al. (2018), Rhinehart et al. (2018a, 2019), Ridel et al. (2019), Ivanovic and Pavone (2019), Amirian et al. (2019), Chai et al. (2019), van der Heiden et al. (2019), Hong et al. (2019), Li et al. (2019), Tang and Salakhutdinov (2019)
	Cumulative probability (CP)	Suraj et al. (2018)

abusing notation, simply MED or ADE. This type of evaluation, however, does not measure how good the predictive distribution matches the ground-truth distribution, falling short of being a true probabilistic measure. For example, it favors point predictions and avoids larger variances, as they often increase the expected ADE.

Dynamic time warping (DTW) (Berndt and Clifford, 1994) computes a similarity metric between trajectories of different length as the minimum total cost of warping one trajectory into another under some distance metric for point pairs. As DTW operates on full trajectories, it is susceptible to outliers.

Modified Hausdorff distance (MHD) (Dubuisson and Jain, 1994) is related to the Hausdorff distance as the maximal minimal distance between the points of predicted and actual trajectory. MHD was designed to be more robust against outliers by allowing slack during matching and to compare trajectories of different length. A further variant is the *trajectory Hausdorff* measure (THAU) (Lee et al., 2007), a path metric that computes a weighted sum over three distance terms each focusing on differences in perpendicular direction, length, and orientation between the paths. The weights can be chosen to be application-dependent.

Longest common subsequence (LCS) (Buzan et al., 2004) aligns two trajectories of different length so as to maximize the length of the common subsequence, i.e., the number of matching points between both trajectories. A good match is determined by thresholding a pair-wise distance and time difference where not all points need to be matched. LCS is more robust to noise and outliers than DTW but finding suitable values for the two thresholds is not always easy.

CLEAR multiple object tracking accuracy (CLEAR-MOTA) was initially introduced as a performance metric for target tracking (Bernardin and Stiefelhagen, 2008). In the context of prediction evaluation, it is similar to LCS in that it sums up good matches between points on the predicted trajectory and the ground truth. The difference is that the concept of pair-wise matches/mismatches is more complex including false negatives, false positives, and non-unique correspondences.

In addition to the metrics considered in Morris and Trivedi (2008) and Quehl et al. (2017), relevant metrics used in the reviewed literature include the *quaternion-based rotationally invariant LCS* (QRLCS), which is the rotationally invariant counterpart of LCS (Hermes et al., 2009), and several measures that quantify different geometric aspects in addition to trajectory or path similarity as follows.

Final displacement error (FDE) measures the distance between final predicted position and the ground-truth position at the corresponding time point. If the prediction is represented by a distribution, many authors compute expected FDE. FDE, however, is not appropriate when there are multiple possible future positions.

Prediction accuracy (PA) uses a binary function to classify a prediction as correct if the predicted position fulfills some criteria, e.g., is within a threshold distance away from the ground truth. Percentage of correctly predicted trajectories is then reported. PA allows to incorporate suitable invariances into the distance function such as allowing certain types of errors.

As also pointed out by Quehl et al. (2017), the challenge in choosing a suitable measure is that each of these measures usually produce quite different results. For the sake of an unbiased and fair evaluation of different prediction algorithms, measures should be chosen not to suit a particular method but based on the requirements from the targeted application. An application that includes a lot of different velocities, for example, should not solely rely on path measures.

7.1.2. Probabilistic accuracy metrics. One of the drawbacks of geometric metrics is their inability to measure uncertainty and also multimodal nature of predictions, e.g., when the target agent may take different paths to reach the goal, or when an observed partial trajectory matches several previously learned motion patterns. Moreover, owing to the stochasticity of the human behaviors, motion prediction algorithms need to be evaluated on their accuracy to match the underlying probability distribution of human movements. Several probabilistic accuracy metrics can be used for this purpose.

Many variational inference and machine learning algorithms (Bishop, 2006; MacKay and Mac Kay, 2003) use the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) to measure dissimilarity of two distributions, e.g., the unknown probability distribution of human behavior $p(\mathbf{s}_{1:T})$ and the predicted probability distribution $q(\mathbf{s}_{1:T}|\theta)$, with θ being a set of parameters of the chosen prediction model. The KL divergence is computed as $d_{KL}(p||q) \simeq \sum_{\mathbf{s}_{1:T}\in\mathbb{S}}\{-p(\mathbf{s}_{1:T})\log q(\mathbf{s}_{1:T}|\theta)+p(\mathbf{s}_{1:T})\log p(\mathbf{s}_{1:T})\}$ with the space of all trajectories $\mathbb{S}$. Minimizing $d_{KL}(p||q)$ corresponds to maximizing the log-likelihood function for θ under the predicted distribution $q(\mathbf{s}_{1:T}|\theta)$. Different surveyed papers have adopted variants of the KL divergence as accuracy metric for their stochastic predictions.

For example, the *average negative log likelihood* or *average negative log loss* evaluates the negative log likelihood term $\left(\simeq \sum_{\mathbf{s}_{1:T}\in\mathbb{D}}\log q(\mathbf{s}_{1:T}|\theta)\right)$ of d_{KL} from a set of ground-truth demonstrations $\mathbb{D} = \left\{\mathbf{s}^i_{1:T}\right\}_{i=1}^N$ with the total number of demonstrations N. Furthermore, several approaches use the *predicted probability* (PP) metric, $\left(\simeq \sum_{t=1}^T q(\mathbf{s}_t|\theta)\right)$ or its negative logarithm, to calculate the probability of the ground-truth path (i.e., $\mathbf{s}_{1:T}$) on the predicted states distribution. For the above metrics, the computation of the log likelihood depends on the chosen model, its induced graph and the corresponding factorization. Finally, the *cumulative probability* (CP) metric computes the fraction of the predictive distribution that lies within a radius r from the correct position for various values of r.

Several recently introduced metrics follow a sampling approach to evaluate a probability distribution. The *minimum average displacement error* (mADE) metric (Rhinehart et al., 2019; Schöller et al., 2019; Tang and Salakhutdinov, 2019; Thiede and Brahma, 2019; Walker et al., 2016), as well as *variety loss*, *oracle*, *minimum over N*, *best-of-N*, *top n%*, or *minimum mean squared distance* (minMSD), computes Euclidean distance between the ground-truth position of the agent $\mathbf{s}^*_t$ at time t and the closest (or the n% closest) of the K samples from the predicted probability distribution: $\min_k \|\mathbf{s}^*_t - \mathbf{s}^k_t\|$. Similarly, *minimum final displacement error* (mFDE) evaluates only the distribution at the prediction horizon T. Such metrics encourage the predicted distribution to cover multiple modes of the ground-truth distribution, while placing probability mass according to the mode likelihood. An evaluation of the robustness of top 1 versus top n% metrics by Bhattacharyya et al. (2019) has shown that the *top n%* metric produces more stable results.

7.1.3. Other performance metrics. Prediction accuracy is by far the primary performance indicator in the reviewed literature across approaches and application domains. In particular, for long-term prediction methods, authors evaluate accuracy against the prediction horizon (Bahram et al., 2016; Blaiotta, 2019; Choi et al., 2019; Chung and Huang, 2010; Deo and Trivedi, 2018; Galceran et al., 2015; Goldhammer et al., 2014; Hermes et al., 2009; Ikeda et al., 2012; Jacobs et al., 2017; Karasev et al., 2016; Keller and Gavrila, 2014; Lee and Kitani, 2016; Pfeiffer et al., 2018, 2016; Quintero et al., 2014; Radwan et al., 2018; Raipuria et al., 2018; Rehder and Klöden, 2015; Rudenko et al., 2018a,b; Sun et al., 2018; Suraj et al., 2018; Thompson et al., 2009; Vasishta et al., 2018; Wu et al., 2018; Xu et al., 2018). Far fewer authors address other aspects of robustness and investigate the range of conditions under which prediction results remain stable and how they are impacted by different types of perturbations.

Experiments to explore robustness evaluate prediction accuracy as a function of various influences: the length or duration of the observed partial trajectory until prediction (addresses the question of how long the target agent needs to be observed for a good prediction) (Kitani et al., 2012; Lee et al., 2017; Radwan et al., 2018), the size of the training dataset (Huynh and Alaghband, 2019; Suraj et al., 2018; Vasishta et al., 2018; Vasquez et al., 2009), number of agents in the scene (Rhinehart et al., 2019), input data sampling frequency and the amount of sensor noise (Bera et al., 2016), or amount of anomalies in the training trajectories (Han et al., 2019). Several authors report a separate accuracy measurement for the more challenging (e.g., non-linear or anomalous) part of the test set (Fernando et al., 2018; Huynh and Alaghband, 2019; Kooij et al., 2019), or evaluate the model's performance on different classes of behavior, e.g., walking or stopping (Saleh et al., 2018b). Analysis of generalization, overfitting, and input utilization by a neural network, presented by Schöller et al. (2019), makes a good case for robustness evaluation.

Furthermore, to quantify efficiency of a prediction method, some authors relate inference time to the number of agents in the scene (Rudenko et al., 2018a,b; Thompson et al., 2009), and only a few papers provide an analysis of their algorithms' complexity (Best and Fitch, 2015; Chen et al., 2016; Keller and Gavrila, 2014; Rudenko et al., 2018b; Zhao et al., 2019).

7.2. Datasets

In order to evaluate the quality of predictions, predicted states or distributions are usually compared with the ground-truth states using standard datasets of recorded motion. Availability of annotated trajectories, represented with the sequence of states or bounding boxes in the top-down view, sets prediction benchmarking datasets aside from the other popular computer vision datasets, where the ground-truth state of the agent is not available and is difficult to estimate.

Common recording setup includes a video camera with static top-down view of the scene, or ground-based lasers and/or depth sensors, mounted on a static or moving platform. Detected agents in each frame are labeled with unique IDs, and their positions with respect to the global world frame are given as (x, y) coordinates together with the frame time-stamp t, i.e., (id, t, x, y). Often the coordinate vector is augmented with orientation and velocity information. Furthermore, social grouping information, gaze directions, motion mode or maneuver labels, and other contextual cues can be provided. Apart from this specific form of labeling, further requirements to prediction benchmarking datasets include interaction between agents, varying density of agents, presence of non-convex obstacles in the environment, availability of the semantic map, and long continuous observations of the agents.

In Table 2 we review the most popular datasets, used for evaluation in the surveyed literature. Out of many datasets, used for benchmarking by different authors, we picked those used by at least two independent teams, excluding the creators of the dataset. We believe that this is a good indication of the dataset's relevance, which also supports the primary purpose of benchchmarking: comparing performance of different methods on the same dataset. In addition, in Table 3 we include four recent datasets, which do not meet the selection criterion, but cover valuable aspects, missing from the earlier datasets. This includes the first dataset of cyclists trajectories (Pool et al., 2017), the first large-scale dataset of vehicles trajectories (Krajewski et al., 2018), the first dedicated benchmark for human trajectory prediction (Sadeghian et al., 2018), and the first dataset of human motion trajectories with accurate motion capture data (Rudenko et al., 2019).

8. Discussion

There has been great progress in developing advanced prediction techniques over recent years in terms of method diversity, performance, and relevance to an increasing number of application scenarios. In this section, we summarize and discuss the state of the art and pose the three questions initially raised in the introduction: *Are the evaluation techniques to measure prediction performance good enough and follow best practices (Q1)*? This is discussed in Section 8.1 by reviewing the existing benchmarking practices including metrics, experiments, and datasets. *Have all prediction methods arrived on the same performance level and the choice of the modeling approach does not matter anymore (Q2)*? This is discussed in Section 8.2 where we consider the theoretical and demonstrated ability of the different modeling approaches to solve the motion prediction problem by accounting for contextual cues from the environment and the target agent. Finally, *Is motion prediction solved (Q3)*? This is discussed in Section 8.3 by revisiting the requirements from the different application scenarios. Finally, in Section 8.4 we outline open challenges and future research directions.

8.1. Benchmarking

Evaluating the performance of a motion prediction algorithm requires choosing appropriate testing scenarios and accuracy metrics, as well as studying the method's robustness against various variables, such as the number of interacting agents or amount of maneuvering in the data.

Depending on the application area, the testing scenario may be an intersection, a highway, a pedestrian crossing, shared urban street with heterogeneous agents, a home environment, or a crowded public space. Existing datasets, summarized in Section 7.2, cover a wide range of scenarios, e.g., indoor (Brščić et al., 2013; Rudenko et al., 2019; Zhou et al., 2012) and outdoor environments (Lerner et al., 2007; Oh et al., 2011; Pellegrini et al., 2009), pedestrian areas (Benfold and Reid, 2011; Majecka, 2009), urban zones (Robicquet et al., 2016; Schneider and Gavrila, 2013) and

Table 2. Overview of the motion trajectories datasets.

Dataset	Location	Agents	Sensors	Scene description	Duration and tracks	Annotations and sampling rate
ETH (Pellegrini et al., 2009)	Outdoor	People	Camera	2 pedestrian scenes, top-down view, moderately crowded	25 min, 650 tracks	Positions, velocities, groups, maps @2.5 Hz
Used by: Varshneya and Srinivasaraghavan (2017), Bera et al. (2016), Alahi et al. (2016), Vemula et al. (2017), Trautman and Krause (2010), Kim et al. (2015), Yamaguchi et al. (2011), Chung and Huang (2010), Vemula et al. (2018), Radwan et al. (2018), Pfeiffer et al. (2018), Bisagno et al. (2018), Zhang et al. (2019), Zhao et al. (2019), Xue et al. (2018), Sadeghian et al. (2019), Gupta et al. (2018), Huynh and Alaghband (2019), Nikhil and Tran Morris (2018), Xu et al. (2018), Luo and Cai (2019), Pei et al. (2019), Huang et al. (2019), Amirian et al. (2019), Blaiotta (2019), Kosaraju et al. (2019), Ivanovic and Pavone (2019)						
UCY (Lerner et al., 2007)	Outdoor	People	Camera	2 pedestrian scenes (sparsely populated Zara and crowded Students), top-down view	16.5 min, over 700 tracks	Positions, gaze directions –
Used by: Ma et al. (2017), Varshneya and Srinivasaraghavan (2017), Alahi et al. (2016), Bartoli et al. (2018), Best and Fitch (2015), Yamaguchi et al. (2011), Pellegrini et al. (2010), Vemula et al. (2018), Radwan et al. (2018), Bisagno et al. (2018), Zhang et al. (2019), Zhao et al. (2019), Hasan et al. (2018), Xue et al. (2018), Sadeghian et al. (2019), Gupta et al. (2018), Huynh and Alaghband (2019), Nikhil and Tran Morris (2018), Xu et al. (2018), van der Heiden et al. (2019), Pei et al. (2019), Huang et al. (2019), Luo and Cai (2019), Amirian et al. (2019), Blaiotta (2019), Kosaraju et al. (2019), Ivanovic and Pavone (2019)						
Stanford Drone Dataset (Robicquet et al., 2016)	Outdoor	People, cyclists, vehicles	Camera	8 urban scenes, $\sim$900 m^2 each, top-down view, moderately crowded	5 hours, 20,000 tracks	Bounding boxes @30 Hz
Used by: Varshneya and Srinivasaraghavan (2017), Jacobs et al. (2017), Coscia et al. (2018), Zhao et al. (2019), Sadeghian et al. (2019), van der Heiden et al. (2019), Chai et al. (2019), Fernando et al. (2019), Makansi et al. (2019), Eiffert and Sukkarieh (2019), Ridel et al. (2019), Saleh et al. (2019)						
NGSIM (Colyar and Halkias, 2006, 2007)	Outdoor	Vehicles	Camera network	Recording of the US Highway 101 and Interstate 80, road segment length 640 and 500 m	90 min	Local and global positions, velocities, lanes, vehicle type, and parameters, @10 Hz
Used by: Kuefler et al. (2017), Deo and Trivedi (2018), Zhao et al. (2019), Altché and de La Fortelle (2017), Li et al. (2019), Kalayeh et al. (2015), Dai et al. (2019), Ding et al. (2019), Tang and Salakhutdinov (2019)						
Edinburgh (Majecka, 2009)	Outdoor	People	Camera	1 pedestrian scene, top-down view, 12×16 m^2, varying density of people	Several months, 92,000 tracks	Positions @9 Hz
Used by: Previtali et al. (2016), Elfring et al. (2014), Rudenko et al. (2017), Xue et al. (2017), Fernando et al. (2018), Carvalho et al. (2019)						
Grand Central Station Dataset (Zhou et al., 2012)	Indoor	People	Camera	Recording in the crowded New York Grand Central train station	33 minutes	Tracklets @25 Hz
Used by: Su et al. (2017), Xue et al. (2017), Xue et al. (2019), Yi et al. (2016), Xu et al. (2018), Fernando et al. (2018)						
VIRAT (Oh et al., 2011)	Outdoor	People, cars, other vehicles	Camera	16 urban scenes, 20–50° camera view angle towards the ground plane, homographies included	25 hours	Bounding boxes, events (e.g., entering a vehicle or using a facility) @10, 5, and 2 Hz
Used by: Previtali et al. (2016), Vasquez (2016), Kitani et al. (2012), Walker et al. (2014), Xie et al. (2013)						
KITTI (Geiger et al., 2012)	Outdoor	People, cyclists, vehicles	Velodyne, 4 cameras	Recorded around the mid-size city of Karlsruhe (Germany), in rural areas and on highways	21 training sequences and 29 test sequences	3D Positions @10 Hz
Used by: Karasev et al. (2016), Wu et al. (2018), Rhinehart et al. (2018a), Lee et al. (2017), Srikanth et al. (2019)						
Town Center Dataset (Benfold and Reid, 2011)	Outdoor	People	Camera	Pedestrians moving along a moderately crowded street	5 minutes, 230 hand labelled tracks	Bounding boxes @15 Hz
Used by: Ma et al. (2017), Xue et al. (2018), Xue et al. (2019), Hasan et al. (2018)						
ATC (Brščić et al., 2013)	Indoor	People	3D range sensors	Recording in a shopping center, 900 m^2 coverage, varying density of people	92 days, long tracks	Positions, orientations, velocities, gaze directions, @10–30 Hz
Used by: Rudenko et al. (2018a,b), Molina et al. (2018)						

(Continued)

Table 2. Continued

Dataset	Location	Agents	Sensors	Scene description	Duration and tracks	Annotations and sampling rate
Daimler Pedestrian Path Prediction Dataset (Schneider and Gavrila, 2013)	Outdoor	People	Stereo camera	Recording from a moving or standing vehicle, pedestrians are crossing the street, stopping at the curb, starting to move or bending in	68 tracks of pedestrians, 4 sec each	Positions, bounding boxes, stereo images, calibration data @17 Hz
Used by: Schulz and Stiefelhagen (2015), Saleh et al. (2018b, 2019)						
L-CAS (Yan et al., 2017)	Indoor	People	Velodyne	Recording in a university building from a moving or stationary robot	49 minutes	Positions, groups, Velodyne scans @10 Hz
Used by: Sun et al. (2018), Radwan et al. (2018)						

Table 3. Additional motion trajectories datasets.

Dataset	Location	Agents	Sensors	Scene description	Duration and tracks	Annotations and sampling rate
Tsinghua-Daimler Cyclist (Pool et al., 2017)	Outdoor	Cyclists	Stereo camera	Recording from a moving vehicle	134 tracks	Positions, road topology @5 Hz
Used by: Saleh et al. (2018a)						
TrajNet (Sadeghian et al., 2018)	Outdoor	People	Cameras	Superset of datasets, collecting also relevant metrics and visualization tools	Superset of image-plane and world-plane datasets	Bounding boxes and tracklets, datasets recording at different frequencies
Used by: Xue et al. (2019)						
highD Dataset (Krajewski et al., 2018)	Outdoor	Vehicles	Camera	Six different highway locations near Cologne, top-down view, varying densities with light and heavy traffic	Over 110,000 vehicles, 447 driven hours	Positions and additional features, e.g., THW, TTC @25 Hz
THÖR (Rudenko et al., 2019)	Indoor	People	Motion capture	Human-robot navigation study in a university lab	Over 600 person and group trajectories in 60 minutes	Positions, head orientations, gaze directions, groups, map, Velodyne scans @100 Hz

highways (Colyar and Halkias, 2006, 2007; Krajewski et al., 2018), and include trajectories of various agents, such as people, cyclists, and vehicles. However, these datasets are usually semi-automatically annotated and, therefore, only provide incomplete and noisy estimation of the ground-truth positions (owing to annotation artifacts). Furthermore, length of the trajectories is often not sufficient for evaluation in some application domains, where long-term predictions are required. Moreover, the amount of interactions between recorded agents is often limited or disbalanced (very few agents are interacting, ergo misinterpreting such cases is not reflected in the lower benchmark scores). Finally, relevant semantic information about static (i.e., grass, crosswalks, sidewalks, streets) and dynamic (i.e., human attributes such as age, gender, or group affiliation) entities is usually not recorded.

Accuracy metrics, described in Section 7.1, offer a rich choice for benchmarking, ranging from computing geometric distances between points (ADE, FDE) also accounting for temporal misalignments (DTW, MHD), to probabilistic policy likelihood measures (NLL) and sampling-based distribution evaluation (mADE). For long-term forecasts made in topologically non-trivial scenarios, results are usually multi-modal and associated with uncertainty. Performance evaluation of such methods should make use of metrics that account for this, such as negative log-likelihood or log-loss derived from the KLD. Not all authors are currently using such metrics. Even for short-term prediction horizons, for which a large majority of authors use geometric metrics only (AED, FDE), probabilistic metrics are preferable as they better reflect the stochastic nature of human motion and the uncertainties involved from imperfect sensing.

Another issue of benchmarking is related to variations in exact metric formulation and different names used for the same metric, e.g., for the ADE- and likelihood-based metrics, as indicated in Section 7.1. In addition, precision is often evaluated on a single arbitrary prediction horizon. These aspects obstruct comparison of the relative precision of various methods.

Furthermore, very few authors currently address robustness as a relevant issue/topic. This is surprising as prediction needs to be robust against a variety of perturbations when deployed in real systems. Examples includes sensing and detection errors, tracking deficiencies, self-localization uncertainties, or map changes.

8.1.1. On question 1. We conclude that *Q1* is not confirmed. Despite the numerous metrics, datasets, and experiment designs used in individual works, benchmarking prediction algorithms lack a systematic approach with common evaluation practices.

For evaluating prediction quality, researchers should opt for more-complex testing scenarios (which include non-convex obstacles, long trajectories, collision avoidance maneuvers, and non-trivial interactions) and the complete set of metrics (both geometric and probabilistic). It is a good practice to condition the forecast precision on various prediction horizons, observation periods, and the complexity of the scene, e.g., defined by how many interacting agents are tracked simultaneously. Furthermore, perfect sensing, perception, and tracking is not always achieved in real-life operation, and therefore algorithms' performance ideally should be investigated in realistic conditions and supported by robustness experiments, e.g., see Section 7.1.3. Performing proper performance analysis would clarify application potential and effective prediction horizon of many methods.

Similar benchmarking practices should be applied to runtime evaluation. Considering efficiency on embedded CPUs of autonomous systems is important for the algorithm's design and evaluation. To prove applicability in real-life scenarios (e.g., in the pipeline with time-sensitive local and global motion planners), discussion should include formal complexity and runtime analysis, conditioned on the scene complexity and prediction horizon.

For a fair objective comparison of the prediction algorithms, developing a standard benchmark with testing scenarios and metrics is becoming a task of critical importance, e.g., given the rapid growth in published literature (see Figure 4). The first attempt to build such a benchmark, TrajNet, is taken by Sadeghian et al. (2018), with the follow up, TrajNet++, to be released soon. TrajNet is based on selected trajectories from the ETH, UCY, and Stanford Drone Dataset and uses the ADE and FDE evaluation metrics. We encourage more researchers to follow this example and contribute to the unification of benchmarking practices.

8.2. Modeling approaches

With such a wide variety of motion modeling approaches, a natural question arises: which one should be preferred? In this section, we discuss the inherent strengths and limitations of different approaches' classes and the efforts to incorporate various contextual cues. This discussion continues in Section 8.3 with highlighting the specifics of several key tasks in the application domains.

Physics-based approaches are suitable in those situations where the effect of other agents or the static environment, and the agent's motion dynamics can be modeled by an explicit transition function. Many of the physics-based approaches naturally handle joint predictions With the choice of an appropriate transition function, physics-based approaches can be readily applied across multiple environments, without the need for training datasets (some data for parameter estimation is useful, though). The downside of using explicitly designed motion models is that they might not capture well the complexity of the real world. The transition functions tend to lack information regarding the "greater picture," both on the spatial and the temporal scale, leading to solutions that represent local minima ("dead ends"). In practice, this limits the usability of physics-based methods to short prediction horizons and relatively obstacle-free environments. All in all, the existence of fast approximate inference, the applicability across multiple domains under mild conditions, and the interpretability make physics-based approaches a popular option for the collision avoidance of the mobile platforms (e.g., self-driving vehicles, service robots) and the people-tracking applications.

Pattern-based approaches are suitable for environments with complex unknown dynamics (e.g., public areas with rich semantics), and can cope with comparatively large prediction horizons. However, this requires ample data that must be collected for training purposes in a particular type of location or scenario. One further issue is the generalization capability of such learned model, whether it can be transferred to a different site, especially if the map topology changes (cf. service robot in an office where the furniture has been moved). Pattern-based approaches tend to be used in non-safety critical applications, where explainability is less of an issue and where the environment is spatially constrained.

Planning-based approaches work well if goals, that the agents try to accomplish, can be explicitly defined and a map of the environment is available. In these cases, the planning-based approaches tend to generate better long-term predictions than the physics-based techniques and generalize to new environments better than the pattern-based approaches. In general, the runtime of planning-based approaches, based on classical planning algorithms (i.e., Dijkstra (Schrijver, 2012), FMM (Sethian, 1996), optimal sampling-based motion planners (Janson et al., 2018; Karaman and Frazzoli, 2011), value iteration (Littman et al., 1995)) scales exponentially with the number of agents, the size of the environment and the prediction horizon (Russell and Norvig, 2016).

8.2.1. On question 2. In our view, *Q2* is not confirmed. As we have seen, the different modeling approaches have various strengths and weaknesses. Although, in principle, it could be possible to incorporate the same contextual cues, there have been so far insufficient studies to compare prediction performance across modeling approaches. Moreover, different modeling approaches exhibit varying degree of complexity and efficiency in including contextual cues from different categories. Physics-based methods are by their very nature aware of the target agent cues and may be easily extended with other ones (e.g., social-force-based (Helbing and Molnar, 1995) and circular distribution-based (Coscia et al., 2018)). Pattern-based methods can potentially handle all kind of contextual information that is encoded in the collected datasets. Some of them are intrinsically map-aware (Bennewitz et al., 2005; Kucner et al., 2013; Roth et al., 2016). Several others can be extended to include further types of contextual information (e.g., Alahi et al., 2016; Bartoli et al., 2018; Pfeiffer et al., 2018; Trautman and Krause, 2010; Vemula et al., 2018), but such extension may lead to involved learning, data efficiency, and generalization issues (e.g., for the clustering methods (Bennewitz et al., 2005; Chen et al., 2008)). Planning-based approaches are intrinsically map- and obstacle-aware, natural to extend with semantic cues (Kitani et al., 2012; Rhinehart et al., 2018a; Rudenko et al., 2018b; Ziebart et al., 2009). Usually they encode the contextual complexity into an objective/reward function, which may fail to properly incorporate dynamic cues (e.g., changing traffic lights). Therefore, authors have to design specific modifications to include dynamic cues into the prediction algorithm (such as jump Markov processes in Karasev et al. (2016), local adaptations of the predicted trajectory in Rudenko et al. (2018a,b), game-theoretic methods in Ma et al. (2017)). Unlike for the pattern-based approaches, target agents cues are natural to incorporate, e.g., as in Kuderer et al. (2012), Rudenko et al. (2018a), and Ma et al. (2017), as both forward and inverse planning approaches rely on a dynamical model of the agents. Contextual cues-dependent parameters of the planning-based methods (e.g., reward functions for inverse planning and models for forward planning) are trivial and typically easier to learn but inference-wise less efficient for high-dimensional (target) agent states compared with the simple physics-based models.

8.3. Application domains

In Section 8.2, we have shown that all modeling approaches theoretically can handle various contextual cues. However, the question of preferring one approach over the others also depends on the task at hand.

8.3.1. Service robots. Predictors for mobile robots usually estimate the most likely future trajectory of each person in the vicinity of the robot. The usual setup includes cameras, range and depth sensors mounted on the robot, operating on a limited-performance mobile CPU.

Physics-based or pattern-based human interaction models, capable of providing short-term high-confidence predictions (i.e., for 1–2 seconds), are best suited for local motion planning and collision avoidance in the crowd. Methods used to this end should have fast and efficient inference for predicting short-term dynamics of several people around the robot. In the simplest case, even linear velocity projection is sufficient for smoothing the robot's local planning (Bai et al., 2015; Chen et al., 2017). More advanced methods should handle human–human interaction (Alahi et al., 2016; Ferrer and Sanfeliu, 2014; Gupta et al., 2018; Moussaïd et al., 2010; Pellegrini et al., 2009), the influence of robot's presence and actions on human motion (Eiffert and Sukkarieh, 2019; Oli et al., 2013; Rhinehart et al., 2019; Schmerling et al., 2018), and high-level body cues of human motion for disambiguating the immediate intention (Hasan et al., 2018; Kooij et al., 2019; Quintero et al., 2014; Unhelkar et al., 2015). In safety-critical applications, reachability-based methods provide a guarantee on local collision avoidance (Bansal et al., 2019). Furthermore, understanding local motion patterns is useful for compliant and unobstructive navigation (Palmieri et al., 2017; Vintr et al., 2019).

For global path and task planning, on the other hand, long-term multi-hypothesis predictions (i.e., for 15–20 seconds ahead) are desired, posing a considerably more challenging task for the prediction system. Reactivity requirement is relaxed, however understanding dynamic (Bera et al., 2017; Ma et al., 2017) and static contextual cues (Chung and Huang, 2010; Coscia et al., 2018; Kitani et al., 2012; Sun et al., 2018), which influence motion in the long-term perspective, reasoning on the map of the environment (Karasev et al., 2016; Rudenko et al., 2018a) and inferring intentions of observed agents (Best and Fitch, 2015; Rehder et al., 2018; Vasquez, 2016) becomes more important. For both local and global path planning, location-independent methods are best suited for predicting motion in a large variety of environments (Bansal et al., 2019; Fernando et al., 2019; Shi et al., 2019).

In terms of accuracy of the current state-of-the-art methods, experimental evaluations on simpler datasets, such as the ETH and UCY, show an average displacement error of 0.19–0.4 m for 4.8 s prediction horizon (Alahi et al., 2016; Radwan et al., 2018; Vemula et al., 2018; Yamaguchi et al., 2011). Linear velocity projection in these scenarios is estimated at 0.53 m ADE. In more-challenging scenarios of the ATC dataset with obstacles and longer trajectories an average error of 1.4–2 m for 9 s prediction has been reported (Alahi et al., 2016; Rudenko et al., 2018b; Sun et al., 2018).

8.3.2. Self-driving vehicles. The early recognition of maneuvers of road users in canonical traffic scenarios is the subject of much interest in the self-driving vehicles application. Several approaches stop short of motion trajectory prediction (i.e., regression) and consider the problem as action classification, while operating on short image sequences. Sensors are typically on-board the vehicle, although some work involves infrastructure-based sensing (e.g., stationary cameras or laser scanners) that can potentially avoid occlusions and provide more precise object localization.

Most works consider the scenario of the laterally crossing pedestrian, dealing with the question what the latter will do at the curbside: start walking, continue walking, or stop walking (Keller and Gavrila, 2014; Kooij et al., 2019, 2014; Schneider and Gavrila, 2013). Some works enlarge the pedestrian crossing scenario, by allowing some initial pedestrian movement along the boardwalk before crossing (Schneider and Gavrila (2013) performed trajectory prediction, while other approaches were limited to crossing intention recognition, e.g. Schneemann and Heinemann (2016), Köhler et al. (2015), and Fang et al. (2017)). This scenario is safety-critical and crucial for autonomous vehicles to solve with high confidence. Pose and high-level contextual cues of the target agent (Kooij et al., 2019), and the scene context modeling (e.g., location and type of the obstacles (Muench and Gavrila, 2019; Völz et al., 2016), state of the traffic lights (Karasev et al., 2016)) are helpful to improve the crossing trajectory prediction.

As to cyclists, Kooij et al. (2019) considered the scenario of a cyclist moving in the same direction as the ego-vehicle, and possibly bending left into the path of the approaching vehicle. Pool et al. (2017) considered the scenario of a cyclist nearing an intersection with up to five different subsequent road directions. Both involved trajectory prediction.

For predicting motion of both cyclists and vehicles is it important to consider multi-modality and uncertainty of the future motion. Recently many authors have proposed solutions to this end (Chai et al., 2019; Cui et al., 2019; Hong et al., 2019; Zhao et al., 2019). Furthermore, it is important to consider coordination of actions between the vehicles (Rhinehart et al., 2019; Schmerling et al., 2018).

It is difficult to compare the experimental results, as the datasets are varying (different timings of same scenario, different sensors, different metrics). Several works report improvements versus their baselines. For example, Figure 2 of Kooij et al. (2014) shows that during pedestrian stopping, 0.9 and 1.1 m improvements in lateral position prediction can be reached with a context-based SLDS, compared with a simpler context-free SLDS and basic LDS (KF), respectively, for prediction horizons up to 1 s. A live vehicle demo of this system at the ECCV'14 conference in Zurich, showed that the superior prediction of the context-based SLDS could lead to evasive vehicle action being triggered up to 1 s earlier than with the basic LDS.

8.3.3. Surveillance. The classification of goals and behaviors as well as the accurate prediction of human motion is of great importance for surveillance applications such as retail analytics or crowd control. Common setups for these applications use stationary sensors to monitor the environment. Although single-frame-based systems allow us to partially solve some tasks such as perimeter protection, incorporating a sequence of observations and making use of behavior prediction models often improve accuracy in cases of occlusions or measurements with low quality (e.g., noise, bad lighting conditions).

Traffic monitoring and management applications can benefit from long-term prediction models, as they allow us to associate new observations with existing tracks (e.g., Luber et al., 2010; Pellegrini et al., 2009, 2010; Yamaguchi et al., 2011) and to model long-term distributions over possible future positions of each person (Chung and Huang, 2012; Yen et al., 2008). Furthermore, it enables the analysis and control of customer flow in populated areas such as malls and airports, by gathering extensive information on human motion patterns (Ellis et al., 2009; Kim et al., 2011; Tay and Laugier, 2008; Yoo et al., 2016), understanding crowd movement in light and dense scenarios, tracking individuals within them, and making future predictions of individuals or crowds (e.g., crowd density prediction). Often these methods benefit from employing sociological methods, such as understanding of social interaction, behavior analysis, group and crowd mobility modeling (Antonini et al., 2006; Bera et al., 2016; Ma et al., 2017; Zhou et al., 2015).

Furthermore identifying deviation from usual patterns often makes the foundation for anomaly detection methods that go beyond perimeter protection, as they analyze trajectories instead of the pure existence of a pedestrian in a specific region.

In addition, in this application area it is difficult to compare results obtained by different approaches, owing to the diversity of the used datasets and the way the evaluation has been performed (e.g., different prediction horizons). In terms of prediction accuracy, we report the most interesting results obtained in densely crowded environments using mainly image data. In these settings, recent state-of-the-art approaches achieve an average displacement error of 0.08–1.2 m on the ETH, UC, NY Grand Central, Town Center, and TrajNet datasets, and a final displacement error of 0.081–2.44 m, with a prediction horizon that generally goes from 0.8 up to 4.8 s (see Xue et al. (2019, 2017, 2018), Zhou et al. (2015), and Shi et al. (2019), the latter using a proprietary dataset and going up to a prediction horizon of 10 s).

8.3.4. On question 3. As we showed in Sections 8.3.1–8.3.3, requirements to the motion prediction framework strongly depend on the application domain and particular use-case scenarios therein (e.g., vehicle merging versus pedestrian crossing within the intelligent vehicles domain). Therefore, it is not possible to conclude achievement of absolute requirements of any sort. When considering concrete use-cases, industry-driven domains, such as intelligent vehicles, appear to be the most mature in terms of formulated requirements and proposed solutions. For instance, requirements to the prediction horizon and metric accuracy for emergency braking of intelligent vehicles in urban driving scenarios are described in the ISO 15622:2018 standard (ISO, 2018), which defines norms for comfortable acceleration/deceleration rates for vehicles, conditioned on the maximum speed and traffic rules, as well as the distribution of pedestrian speed and acceleration. Therefore, we conclude that for specific use-cases, in particular for basic emergency braking for intelligent vehicles, solutions have achieved a level of performance that allows for industrialization into consumer products. Those use-cases can be considered solved. For other use-cases we expect more standardization and explicit formulation of requirements to take place in the near future. For instance, the standard for safety requirements for personal care robots, Standard ISO 13482:2014 (ISO, 2014) suggests using sensors for detecting a human in the vicinity of the robot to issue a protective stop, and controlling the speed and force when the robot is in close proximity to humans to reduce the risk of collision. This standard, however, does not propose motion anticipation to improve the risk assessment.

Furthermore, several aspects of performance, robustness, and generalization to new environments, discussed in the following sections, need to be explored before reaching further conclusions on maturity of the solutions. Finally, in order to reliably assess the quality of existing solutions across all application domains, is it critical to address the issues of benchmarking.

8.4. Future directions

Developing more sophisticated methods for motion prediction which go beyond Kalman filtering with simple motion models is a clear trend of recent years. Modern techniques make extensive use of machine learning in order to better estimate context-dependent patterns in real data, handle more complex environment models and types of motion, or even propose end-to-end reasoning on future motion from visual input. An increasing number of methods also includes reasoning on the global structure of the environment, intentions, and actions of the agent. Having these trends in mind, we see several directions of future research.

8.4.1. Use of enhanced contextual cues. To analyze and predict human motion, as well as to plan and navigate alongside them, intelligent systems should have an in-depth semantic scene understanding. Context understanding with respect to features of the static environment and its semantics for better trajectory prediction is still a relatively unexplored area, see Section 6.3 for more details.

The same argument applies for the contextual cues of the dynamic environment. Socially aware methods are making an important improvement over socially unaware ones in such spaces where the target agent is not acting in isolation. However, most existing socially aware methods still assume that all observed people are behaving similarly and that their motion can be predicted by the same model and with the same features. Capturing and reasoning on the high-level social attributes is at an early stage of development, see Sections 6.1 and 6.2, however recent methods take steps. Furthermore, most available approaches assume cooperative behavior, while real humans might rather optimize personal goals instead of joint strategies. In such cases, game-theoretic approaches are possibly better suited for modeling human behavior. Consequently, adopting classical artificial intelligence and game-theoretic approaches in multi-agent systems is a promising research direction, that is only partly addressed in recent work (see, e.g., Bahram et al., 2016; Ma et al., 2017).

One task where contextual cues become particularly important is long-term prediction of motion trajectories. While context-agnostic motion and behavioral patterns are helpful for short prediction horizons, long-term predictions should account for intentions, based on the context and the surrounding environment. Many pattern-based methods treat agents as particles, placed in the field of learned transitions, dictating the direction of future motion. Extending these models by more goal- or intention-driven predictions, that resemble human goal-directed behavior, would be beneficial for long-term predictions.

Consequently, further research on automatic goal inference based on the semantics of the environment is important. Most planning-based methods rely on a given set of goals, which makes them unusable or imprecise in a situation where no goals are known beforehand, or the number of possible goals is too high. Alternatively, one could consider identifying on-the-fly possible goals in the environment and predicting the way the agent may reach those goals. This would allow application of the planning-based methods in unknown environments. In addition, semantic indicators of possible goals, coming from understanding the person's social role or current activity (Bruckschen et al., 2019), could lead to more robust intention recognition.

Apart from the contextual cues, discussed in this survey, there are many other factors influencing pedestrian motion, according to the recent studies (Rasouli and Tsotsos, 2019), e.g., weather conditions, time of day, social roles of agents. Future methods could benefit from closer connection to the studies of human motion and behavior in social spaces (Arechavaleta et al., 2008; Do et al., 2016; Gorrini et al., 2016).

8.4.2. Robustness and integration. Several practical aspects of deploying prediction systems in real environments should be considered in the future work.

Most of the presented methods are designed for specific tasks, scenarios, or types of motion. These methods work well in certain situations, e.g., when prominent motion patterns exist in the environment, or when the spatial structure of the environment and target agent's goals are known beforehand. A conceptually interesting approach that uses a combination of multiple prediction algorithms to reason about best performance in the given situation is presented by Lasota and Shah (2017). The multiple-predictor framework opens a possibility for achieving more robust predictions when operating in undefined, changing situations, where a combination of strengths of different methods is required.

We suggest that more emphasis should be put on transfer learning and generalization of approaches to new environments. Learning and reasoning on basic, invariant rules and norms of human motion and collision avoidance is a better approach in this case. When having access to several environments, domain adaptation could be potentially used for learning generalizable models.

Integration of prediction in planning and control is another worthwhile topic for overall system robustness. Predicting human motion is usually motivated with increased safety of human–robot interaction and efficiency of operation. However, the insights on exploiting predictions in the robot's motion or action planning module are typically left out of scope in many papers. Future work would benefit from outlining possible ways to incorporate predictions in the robot control framework.

9. Conclusions

In this work, we have presented a thorough analysis of the human motion trajectory prediction problem. We have surveyed the literature across multiple domains and proposed a taxonomy of motion prediction techniques. Our taxonomy builds on the two fundamental aspects of the motion prediction problem: the model of motion and the input contextual cues. We have reviewed the relevant trajectory prediction tasks in several application areas, such as service robotics, self-driving vehicles, and advance surveillance systems. Finally, we summarize and discuss the state of the art along the lines of three major questions and outlined several prospective directions of future research.

"*Prediction is very difficult, especially about the future*." This quote (whose origin has been attributed to multiple people) certainly remains applicable to motion trajectory prediction, despite two decades of research and the >200 prediction methods listed in this survey. We hope that our survey increases visibility in this rapidly expanding field and the will stimulate further research along the directions discussed.

Acknowledgement

The authors would like to thank Achim J Lilienthal for valuable feedback and suggestions.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work has been partly funded from the European Union's Horizon 2020 research and innovation programme (grant agreement number 732737 (ILIAD)).

ORCID iDs

Andrey Rudenko http://orcid.org/0000-0002-2181-3444
Luigi Palmieri http://orcid.org/0000-0002-4908-5434

References

Agamennoni G, Nieto JI and Nebot EM (2012) Estimation of multivehicle dynamics by considering contextual information. *IEEE Transactions on Robotics* 28(4): 855–870.

Alahi A, Goel K, Ramanathan V, Robicquet A, Fei-Fei L and Savarese S (2016) Social LSTM: Human trajectory prediction in crowded spaces. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–971.

Altché F and de La Fortelle A (2017) An LSTM network for highway trajectory prediction. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*. IEEE, pp. 353–359.

Althoff M (2010) *Reachability Analysis and Its Application to the Safety Assessment of Autonomous Cars*. PhD Thesis, TU Munich.

Althoff M, Heß D and Gambert F (2013) Road occupancy prediction of traffic participants. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 99–105.

Althoff M, Stursberg O and Buss M (2008a) Reachability analysis of nonlinear systems with uncertain parameters using conservative linearization. In: *Proceedings of the IEEE International Conference on Decision and Control (CDC)*, pp. 4042–4048.

Althoff M, Stursberg O and Buss M (2008b) Stochastic reachable sets of interacting traffic participants. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 1086–1092.

Amirian J, Hayet JB and Pettré J (2019) Social ways: Learning multi-modal distributions of pedestrian trajectories with GANs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Antonini G, Martinez SV, Bierlaire M and Thiran JP (2006) Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision* 69(2): 159–180.

Aoude G, Joseph J, Roy N and How J (2011) Mobile agent trajectory prediction using Bayesian nonparametric reachability trees. In: *Proceedings of AIAA Infotech@Aerospace (I@A)*, pp. 1–17.

Aoude GS, Luders BD, Lee KKH, Levine DS and How JP (2010) Threat assessment design for driver assistance system at intersections. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 1855–1862.

Arechavaleta G, Laumond JP, Hicheur H and Berthoz A (2008) An optimality principle governing human walking. *IEEE Transactions on Robotics and Automation* 24(1): 5–14.

Bahram M, Lawitzky A, Friedrichs J, Aeberhard M and Wollherr D (2016) A game-theoretic approach to replanning-aware interactive scene prediction and planning. *IEEE Transactions on Vehicle Technology* 65(6): 3981–3992.

Bai H, Cai S, Ye N, Hsu D and Lee WS (2015) Intention-aware online POMDP planning for autonomous driving in a crowd. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 454–460.

Ballan L, Castaldo F, Alahi A, Palmieri F and Savarese S (2016) Knowledge transfer for scene-specific motion prediction. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 697–713.

Bandyopadhyay T, Won KS, Frazzoli E, Hsu D, Lee WS and Rus D (2013) Intention-aware motion planning. In: *Algorithmic Foundations of Robotics X*. Berlin: Springer, pp. 475–491.

Bansal S, Bajcsy A, Ratner E, Dragan AD and Tomlin CJ (2019) A Hamilton–Jacobi reachability-based framework for predicting and analyzing human motion for safe planning. *arXiv preprint arXiv:1910.13369*.

Barth A and Franke U (2008) Where will the oncoming vehicle be the next second? In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 1068–1073.

Bartoli F, Lisanti G, Ballan L and Bimbo AD (2018) Context-aware trajectory prediction. In: *Proceedings of the IEEE International Conference on Pattern Recognition*, pp. 1941–1946.

Batkovic I, Zanon M, Lubbe N and Falcone P (2018) A computationally efficient model for pedestrian motion prediction In: *2018 European Control Conference (ECC)*, pp. 374–379.

Batz T, Watson K and Beyerer J (2009) Recognition of dangerous situations within a cooperative group of vehicles. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 907–912.

Benfold B and Reid I (2011) Stable multi-target tracking in real-time surveillance video. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3457–3464.

Bennewitz M, Burgard W, Cielniak G and Thrun S (2005) Learning motion patterns of people for compliant robot motion. *The International Journal of Robotics Research* 24(1): 31–48.

Bennewitz M, Burgard W and Thrun S (2002) Using EM to learn motion behaviors of persons with mobile robots. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 502–507.

Bera A, Kim S, Randhavane T, Pratapa S and Manocha D (2016) GLMP-realtime pedestrian path prediction using global and local movement patterns. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5528–5535.

Bera A, Randhavane T and Manocha D (2017) Aggressive, tense, or shy? Identifying personality traits from crowd videos. In: *Proceedings of the International Conference on Artificial Intelligence (IJCAI)*, pp. 112–118.

Bernardin K and Stiefelhagen R (2008) Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing* 2008(1): 246309.

Berndt DJ and Clifford J (1994) Using dynamic time warping to find patterns in time series. In: *Workshop Proceedings of the AAAI Conference on Artificial Intelligence on Knowledge Discovery and Data Mining (AAAIWS'94)*. AAAI Press, pp. 359–370.

Best G and Fitch R (2015) Bayesian intention inference for trajectory prediction with an unknown goal destination. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 5817–5823.

Best R and Norton J (1997) A new model and efficient tracker for a target with curvilinear motion. *IEEE Transactions on Aerospace and Electronic Systems* 33(3): 1030–1037.

Bhattacharya S, Kumar V and Likhachev M (2010) Search-based path planning with homotopy class constraints. In: *Proceedings of the Annual Symposium on Combined Search*.

Bhattacharyya A, Hanselmann M, Fritz M, Schiele B and Straehle CN (2019) Conditional flow variational autoencoders for structured sequence prediction. *arXiv preprint arXiv: 1908.09008*.

Bisagno N, Zhang B and Conci N (2018) Group LSTM: Group trajectory prediction in crowded scenarios. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 213–225.

Bishop CM (2006) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin: Springer-Verlag.

Blaiotta C (2019) Learning generative socially aware models of pedestrian motion. *IEEE Robotics and Automation Letters* 4(4): 3433–3440.

Brewer MA, Fitzpatrick K, Whitacre JA and Lord D (2006) Exploration of pedestrian gap-acceptance behavior at selected locations. *Transportation Research Record* 1982(1): 132–140.

Broadhurst A, Baker S and Kanade T (2005) Monte Carlo road safety reasoning. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 319–324.

Brouwer N, Kloeden H and Stiller C (2016) Comparison and evaluation of pedestrian motion models for vehicle safety systems. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 2207–2212.

Brown GW (1951) Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation* 13(1): 374–376.

Bruce A and Gordon G (2004) Better motion prediction for people-tracking. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Bruckschen L, Dengler N and Bennewitz M (2019) Human motion prediction based on object interactions. In: *Proceedings of the European Conference on Mobile Robots (ECMR)*. IEEE, pp. 1–6.

Brščić D, Kanda T, Ikeda T and Miyashita T (2013) Person tracking in large public spaces using 3-D range sensors. *IEEE Transactions on Human–Machine Systems* 43(6): 522–534.

Buzan D, Sclaroff S and Kollios G (2004) Extraction and clustering of motion trajectories in video. In: *Proceedings of the IEEE International Conference on Pattern Recognition*, Vol. 2, pp. 521–524.

Cai Y, de Freitas N and Little JJ (2006) Robust visual tracking for multiple targets. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 107–118.

Carvalho JF, Vejdemo-Johansson M, Pokorny FT and Kragic D (2019) Long-term prediction of motion trajectories using path homology clusters. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Chai Y, Sapp B, Bansal M and Anguelov D (2019) Multi-Path: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*.

Chen YF, Liu M, Everett M and How JP (2017) Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 285–292.

Chen YF, Liu M and How JP (2016) Augmented dictionary learning for motion prediction. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2527–2534.

Chen Z, Ngai DCK and Yung NHC (2008) Pedestrian behavior prediction based on motion patterns for vehicle-to-pedestrian collision avoidance. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 316–321.

Chik SF, Yeong CF, Su ELM, Lim TY, Subramaniam Y and Chin PJH (2016) A review of social-aware navigation frameworks for service robot in dynamic human environments. *Journal of Telecommunications, Electronic and Computer Engineering* 8(11): 41–50.

Choi C, Patil A and Malla S (2019) Drogon: A causal reasoning framework for future trajectory forecast. *arXiv preprint arXiv:1908.00024*.

Choi W and Savarese S (2010) Multiple target tracking in world coordinate with single, minimally calibrated camera. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 553–567.

Chung SY and Huang HP (2010) A mobile robot that understands pedestrian spatial behaviors. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 5861–5866.

Chung SY and Huang HP (2012) Incremental learning of human social behaviors with feature-based spatial effects. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 2417–2422.

Colyar J and Halkias J (2006) US highway 80 dataset. *Technical Report*, Federal Highway Administration (FHWA).

Colyar J and Halkias J (2007) US highway 101 dataset. *Technical Report FHWA-HRT-07-030*, Federal Highway Administration (FHWA).

Coscia P, Castaldo F, Palmieri FAN, Alahi A, Savarese S and Ballan L (2018) Long-term path prediction in urban scenarios using circular distributions. *Image and Vision Computing* 69: 81–91.

Cui H, Radosavljevic V, Chou FC, et al. (2019) Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Dai S, Li L and Li Z (2019) Modeling vehicle interactions via modified LSTM models for trajectory prediction. *IEEE Access* 7: 38287–38296.

Deo N and Trivedi MM (2018) Multi-modal trajectory prediction of surrounding vehicles with maneuver based LSTMs. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 1179–1184.

Ding W, Chen J and Shen S (2019) Predicting vehicle behaviors over an extended horizon using behavior interaction network. *arXiv preprint arXiv:1903.00848*.

Djuric N, Radosavljevic V, Cui H, et al. (2018) Motion prediction of traffic actors for autonomous driving using deep convolutional networks. *arXiv preprint arXiv:1808.05819*.

Do T, Haghani M and Sarvi M (2016) Group and single pedestrian behavior in crowd dynamics. *Transportation Research Record* 2540(1): 13–19.

Dubuisson MP and Jain AK (1994) A modified Hausdorff distance for object matching. In: *Proceedings of the IEEE International Conference on Pattern Recognition*, Vol. 1. IEEE, pp. 566–568.

Eiffert S and Sukkarieh S (2019) Predicting responses to a robot's future motion using generative recurrent neural networks. *arXiv preprint arXiv:1909.13486*.

Elfring J, van de Molengraft R and Steinbuch M (2014) Learning intentions for improved human motion prediction. *Journal of Robotics and Autonomous Systems* 62(4): 591–602.

Ellis D, Sommerlade E and Reid I (2009) Modelling pedestrian trajectory patterns with gaussian processes. In: *Proceedings of the International Conference on Computer Vision Workshop*. IEEE, pp. 1229–1234.

Elnagar A (2001) Prediction of moving objects in dynamic environments using Kalman filters. In: *Proceedings of the IEEE International Symposium on Computer Intelligence in Robotics and Automation (CIRA)*, pp. 414–419.

Elnagar A and Gupta K (1998) Motion prediction of moving objects based on autoregressive model. *IEEE Transactions on Systems, Man, and Cybernetics (SMC) - Part A: Systems and Humans* 28(6): 803–810.

Fang Z, Vázquez D and López AM (2017) On-board detection of pedestrian intentions. *Sensors* 17(10): 2193.

Farina F, Fontanelli D, Garulli A, Giannitrapani A and Prattichizzo D (2017) Walking ahead: The headed social force model. *PloS ONE* 12(1): e0169734.

Fearnhead P and Liu Z (2007) On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(4): 589–605.

Ferguson S, Luders B, Grande RC and How JP (2015) Real-time predictive modeling and robust avoidance of pedestrians with uncertain, changing intentions. In: *Algorithmic Foundations of Robotics XI*. Berlin: Springer, pp. 161–177.

Fernando T, Denman S, Sridharan S and Fookes C (2018) Soft+hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection. *Neural Networks* 108: 466–478.

Fernando T, Denman S, Sridharan S and Fookes C (2019) Neighbourhood context embeddings in deep inverse reinforcement learning for predicting pedestrian motion over long time horizons. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*.

Ferrer G and Sanfeliu A (2014) Behavior estimation for a complete framework for human motion prediction in crowded environments. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5940–5945.

Foka AF and Trahanias PE (2002) Predictive autonomous robot navigation. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Foka AF and Trahanias PE (2010) Probabilistic autonomous robot navigation in dynamic environments with human motion prediction. *International Journal of Social Robotics* 2(1): 79–94.

Galceran E, Cunningham AG, Eustice RM and Olson E (2015) Multipolicy decision-making for autonomous driving via changepoint-based behavior prediction. In: *Proceedings of the Robotics: Science and Systems (RSS)*.

Geiger A, Lenz P and Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gindele T, Brechtel S and Dillmann R (2010) A probabilistic model for estimating driver behaviors and vehicle trajectories in traffic environments. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 1625–1631.

Goldhammer M, Doll K, Brunsmann U, Gensler A and Sick B (2014) Pedestrian's trajectory forecast in public traffic with artificial neural networks. In: *Proceedings of the IEEE International Conference on Pattern Recognition*, pp. 4110–4115.

Gong H, Sim J, Likhachev M and Shi J (2011) Multi-hypothesis motion planning for visual object tracking. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 619–626.

Goodfellow I, Pouget-Abadie J, Mirza M, et al. (2014) Generative adversarial nets. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680.

Gorrini A, Vizzari G and Bandini S (2016) Age and group-driven pedestrian behaviour: From observations to simulations. *Collective Dynamics* 1: 1–16.

Gu Y, Hashimoto Y, Hsu LT and Kamijo S (2016) Motion planning based on learning models of pedestrian and driver behaviors. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 808–813.

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

Habibi G, Jaipuria N and How JP (2018) Context-aware pedestrian motion prediction in urban intersections. *arXiv preprint arXiv:1806.09453*.

Han Y, Tse R and Campbell M (2019) Pedestrian motion model using non-parametric trajectory clustering and discrete transition points. *IEEE Robotics and Automation Letters* 4(3): 2614–2621.

Hasan I, Setti F, Tsesmelis T, Del Bue A, Galasso F and Cristani M (2018) MX-LSTM: Mixing tracklets and vislets to jointly forecast trajectories and head poses. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6067–6076.

Helbing D and Molnar P (1995) Social force model for pedestrian dynamics. *Physical Review E* 51(5): 4282.

Henry P, Vollmer C, Ferris B and Fox D (2010) Learning to navigate through crowded environments. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 981–986.

Hermes C, Wöhler C, Schenk K and Kummert F (2009) Long-term vehicle motion prediction. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 652–657.

Hirakawa T, Yamashita T, Tamaki T and Fujiyoshi H (2018) Survey on vision-based path prediction. In: *Distributed, Ambient and Pervasive Interactions: Technologies and Contexts*. Springer International Publishing, pp. 48–64.

Ho J and Ermon S (2016) Generative adversarial imitation learning. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 4565–4573.

Hoermann S, Stumper D and Dietmayer K (2017) Probabilistic long-term prediction for autonomous vehicles. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 237–243.

Hofbaur MW and Williams BC (2004) Hybrid estimation of complex systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 34(5): 2178–2191.

Hong J, Sapp B and Philbin J (2019) Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8454–8462.

Huang S, Li X, Zhang Z, et al. (2016) Deep learning driven visual path prediction from a single image. *IEEE Transactions on Image Processing* 25(12): 5892–5904.

Huang Y, Bi H, Li Z, Mao T and Wang Z (2019) STGAT: Modeling spatial–temporal interactions for human trajectory prediction. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 6272–6281.

Huynh M and Alaghband G (2019) Trajectory prediction by coupling scene-LSTM with human movement LSTM. In: *International Symposium on Visual Computing*. Berlin: Springer, pp. 244–259.

Ikeda T, Chigodo Y, Rea D, Zanlungo F, Shiomi M and Kanda T (2012) Modeling and prediction of pedestrian behavior based on the sub-goal concept. *Proceedings of the Robotics: Science and Systems (RSS)*.

ISO 13482:2014 (2014) Robots and robotic devices – Safety requirements for personal care robots. Standard ISO 13482:2014. International Standards Organization.

ISO (2018) Intelligent transport systems – Adaptive cruise control systems – Performance requirements and test procedures. Standard ISO 15622:2018. International Standards Organization.

Ivanovic B and Pavone M (2019) The Trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2375–2384.

Jacobs HO, Hughes OK, Johnson-Roberson M and Vasudevan R (2017) Real-time certified probabilistic pedestrian forecasting. *IEEE Robotics and Automation Letters* 2(4): 2064–2071.

Jain A, Casas S, Liao R, et al. (2019) Discrete residual flow for probabilistic pedestrian behavior prediction. *arXiv preprint arXiv:1910.08041*.

Jain A, Zamir AR, Savarese S and Saxena A (2016) Structural-RNN: Deep learning on spatio-temporal graphs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5308–5317.

Janson L, Ichter B and Pavone M (2018) Deterministic sampling-based motion planning: Optimality, complexity, and performance. *The International Journal of Robotics Research* 37(1): 46–61.

Joseph J, Doshi-Velez F, Huang AS and Roy N (2011) A Bayesian nonparametric approach to modeling motion patterns. *Journal of Autonomous Robots* 31(4): 383.

Kaempchen N, Weiss K, Schaefer M and Dietmayer KCJ (2004) IMM object tracking for high dynamic driving maneuvers. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 825–830.

Käfer E, Hermes C, Wöhler C, Ritter H and Kummert F (2010) Recognition of situation classes at road intersections. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3960–3965.

Kalayeh MM, Mussmann S, Petrakova A, Lobo NdV and Shah M (2015) Understanding trajectory behavior: A motion pattern approach. *arXiv preprint arXiv:1501.00614*.

Karaman S and Frazzoli E (2011) Sampling-based algorithms for optimal motion planning. *The International Journal of Robotics Research* 30(7): 846–894.

Karamouzas I, Heil P, van Beek P and Overmars MH (2009) A predictive collision avoidance model for pedestrian simulation. In: *International Workshop on Motion in Games*. Berlin: Springer, pp. 41–52.

Karamouzas I and Overmars M (2010) A velocity-based approach for simulating human collision avoidance. In: *Proceedings of the International Conference on Intelligent Virtual Agents*. Berlin: Springer, pp. 180–186.

Karamouzas I and Overmars M (2012) Simulating and evaluating the local behavior of small pedestrian groups. *IEEE Transactions on Visualization and Computer Graphics* 18(3).

Karasev V, Ayvaci A, Heisele B and Soatto S (2016) Intent-aware long-term prediction of pedestrian motion. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2543–2549.

Keller CG and Gavrila DM (2014) Will the pedestrian cross? A study on pedestrian path prediction. *IEEE Transactions on Intelligent Transport Systems* 15(2): 494–506.

Keller CG, Hermes C and Gavrila DM (2011) Will the pedestrian cross? Probabilistic path prediction based on learned motion features. In: *Pattern Recognition: DAGM Symposium*, pp. 386–395.

Kim B, Kang CM, Kim J, Lee SH, Chung CC and Choi JW (2017) Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 399–404.

Kim K, Lee D and Essa I (2011) Gaussian process regression flow for analysis of motion trajectories. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1164–1171.

Kim S, Guy SJ, Liu W, et al. (2015) BRVO: Predicting pedestrian trajectories using velocity-space reasoning. *The International Journal of Robotics Research* 34(2): 201–217.

Kirubarajan T, Bar-Shalom Y, Pattipati KR and Kadar I (2000) Ground target tracking with variable structure IMM estimator. *IEEE Transactions on Aerospace and Electronic Systems* 36(1): 26–46.

Kitani KM, Ziebart BD, Bagnell JA and Hebert M (2012) Activity forecasting. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 201–214.

Köhler S, Goldhammer M, Zindler K, Doll K and Dietmeyer K (2015) Stereo-vision-based pedestrian's intention detection in a moving vehicle. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 2317–2322.

Koller D, Friedman N and Bach F (2009) *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Kooij JFP, Flohr F, Pool EAI and Gavrila DM (2019) Context-based path prediction for targets with switching dynamics. *International Journal of Computer Vision* 127(3): 239–262.

Kooij JFP, Schneider N, Flohr F and Gavrila DM (2014) Context-based pedestrian path prediction. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 618–633.

Kosaraju V, Sadeghian A, Martín-Martín R, Reid I, Rezatofighi SH and Savarese S (2019) Social-bigat: Multimodal trajectory forecasting using Bicycle-GAN and graph attention networks. *arXiv preprint arXiv:1907.03395*.

Koschi M, Pek C, Beikirch M and Althoff M (2018) Set-based prediction of pedestrians in urban environments considering formalized traffic rules. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*.

Krajewski R, Bock J, Kloeker L and Eckstein L (2018) The highD dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*.

Kretzschmar H, Kuderer M and Burgard W (2014) Learning to predict trajectories of cooperatively navigating agents. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4015–4020.

Kruse E and Wahl FM (1998) Camera-based observation of obstacle motions to derive statistical data for mobile robot motion planning. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Vol. 1, pp. 662–667.

Kruse T, Pandey AK, Alami R and Kirsch A (2013) Human-aware robot navigation: A survey. *Journal of Robotics and Autonomous Systems* 61(12): 1726–1743.

Kucner TP, Magnusson M, Schaffernicht E, Bennetts VH and Lilienthal AJ (2017) Enabling flow awareness for mobile robots in partially observable environments. *IEEE Robotics and Automation Letters* 2(2): 1093–1100.

Kucner TP, Saarinen J, Magnusson M and Lilienthal AJ (2013) Conditional transition maps: Learning motion patterns in dynamic environments. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 1196–1201.

Kuderer M, Kretzschmar H, Sprunk C and Burgard W (2012) Feature-based prediction of trajectories for socially compliant navigation. In: *Proceedings of the International Conference on Robotics: Science and Systems*.

Kuefler A, Morton J, Wheeler T and Kochenderfer M (2017) Imitating driver behavior with generative adversarial networks. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 204–211.

Kuhnt F, Kohlhaas R, Schamm T and Zöllner JM (2015) Towards a unified traffic situation estimation model-street-dependent behaviour and motion models. In: *Proceedings of the IEEE International Conference on Information Fusion (Fusion)*, pp. 1223–1229.

Kuhnt F, Schulz J, Schamm T and Zöllner JM (2016) Understanding interactions between traffic participants based on learned behaviors. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 1271–1278.

Kullback S and Leibler RA (1951) On information and sufficiency. *The Annals of Mathematical Statistics* 22(1): 79–86.

Kuwata Y, Teo J, Fiore G, Karaman S, Frazzoli E and How JP (2009) Real-time motion planning with applications to autonomous urban driving. *IEEE Transactions on Control Systems Technology* 17(5): 1105–1118.

Lasota PA, Fong T and Shah JA (2017) A survey of methods for safe human–robot interaction. *Foundations and Trends in Robotics* 5(4): 261–349.

Lasota PA and Shah JA (2017) A multiple-predictor approach to human motion prediction. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2300–2307.

Lee JG, Han J and Whang KY (2007) Trajectory clustering: A partition-and-group framework. In: *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data (SIGMOD '07)*. New York: ACM Press, pp. 593–604.

Lee N, Choi W, Vernaza P, Choy CB, Torr PHS and Chandraker M (2017) DESIRE: Distant future prediction in dynamic scenes with interacting agents. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 336–345.

Lee N and Kitani KM (2016) Predicting wide receiver trajectories in american football. In: *Proceedings of the IEEE Winter*

Conference on Applications of Computer Vision (WACV), pp. 1–9.

Lefèvre S, Vasquez D and Laugier C (2014) A survey on motion prediction and risk assessment for intelligent vehicles. *Robomech Journal* 1: 1.

Lerner A, Chrysanthou Y and Lischinski D (2007) Crowds by example. *Computer Graphics Forum* 26: 655–664.

Li J, Ma H, Zhan W and Tomizuka M (2019) Coordination and trajectory prediction for vehicle interactions via bayesian generative modeling. *arXiv preprint arXiv:1905.00587*.

Li XR and Jilkov VP (2003) Survey of maneuvering target tracking. Part I: Dynamic models. *IEEE Transactions on Aerospace and Electronic Systems* 39(4): 1333–1364.

Li XR and Jilkov VP (2005) Survey of maneuvering target tracking. Part V: Multiple-model methods. *IEEE Transactions on Aerospace and Electronic Systems* 41(4): 1255–1321.

Li XR and Jilkov VP (2010) Survey of maneuvering target tracking. part II: Motion models of ballistic and space targets. *IEEE Transactions on Aerospace and Electronic Systems* 46(1): 96–119.

Li Y, Song J and Ermon A (2017) Infogail: Interpretable imitation learning from visual demonstrations. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 3812–3822.

Liao L, Fox D, Hightower J, Kautz H and Schulz D (2003) Voronoi tracking: Location estimation using sparse and noisy sensor data. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, Vol. 1, pp. 723–728.

Liebner M, Klanner F, Baumann M, Ruhhammer C and Stiller C (2013) Velocity-based driver intent inference at urban intersections in the presence of preceding vehicles. *IEEE Intelligent Transport Systems Magazine* 5(2): 10–21.

Littman ML, Dean TL and Kaelbling LP (1995) On the complexity of solving Markov decision problems. In: *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. San Mateo, CA: Morgan Kaufmann, pp. 394–402.

Luber M, Spinello L, Silva J and Arras KO (2012) Socially-aware robot navigation: A learning approach. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 902–907.

Luber M, Stork JA, Tipaldi GD and Arras KO (2010) People tracking with human motion predictions from social forces. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 464–469.

Luber M, Tipaldi GD and Arras KO (2011) Place-dependent people tracking. *The International Journal of Robotics Research* 30(3): 280–293.

Luo Y and Cai P (2019) GAMMA: A general agent motion prediction model for autonomous driving. *arXiv preprint arXiv:1906.01566*.

Ma WC, Huang DA, Lee N and Kitani KM (2017) Forecasting interactive dynamics of pedestrians with fictitious play. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4636–4644.

MacKay DJ and Mac Kay DJ (2003) *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

Majecka B (2009) Statistical models of pedestrian behaviour in the forum. *Master's Thesis*, School of Informatics, University of Edinburgh.

Makansi O, Ilg E, Cicek O and Brox T (2019) Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7144–7153.

Makris D and Ellis T (2002) Path detection in video surveillance. *Image and Vision Computing* 20(12): 895–903.

Mavrogiannis CI and Knepper RA (2016) Decentralized multi-agent navigation planning with braids. In: *Proceedings of the 2016 International Workshop on the Algorithmic Foundations of Robotics (WAFR'16)*.

Mazor E, Averbuch A, Bar-Shalom Y and Dayan J (1998) Interacting multiple model methods in target tracking: A survey. *IEEE Transactions on Aerospace and Electronic Systems* 34(1): 103–123.

Mínguez RQ, Alonso IP, Fernández-Llorca D and Sotelo MÁ (2018) Pedestrian path, pose, and intention prediction through Gaussian process dynamical models and pedestrian activity recognition. *IEEE Transactions on Intelligent Transport Systems*, in press.

Møgelmose A, Trivedi MM and Moeslund TB (2015) Trajectory analysis and prediction for improved pedestrian safety: Integrated framework and evaluations. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 330–335.

Molina S, Cielniak G, Krajník T and Duckett T (2018) Modelling and predicting rhythmic flow patterns in dynamic environments. In: *Annual Conference Towards Autonomous Robotics Systems*. Berlin: Springer, pp. 135–146.

Morris BT and Trivedi MM (2008) A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Transactions on Circuits and Systems for Video Technology* 18(8): 1114–1127.

Moussaïd M, Perozo N, Garnier S, Helbing D and Theraulaz G (2010) The walking behaviour of pedestrian social groups and its impact on crowd dynamics. *PloS ONE* 5(4): e10047.

Muench C and Gavrila DM (2019) Composable Q-functions for pedestrian car interactions. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*. IEEE.

Murino V, Cristani M, Shah S and Savarese S (2017) *Group and Crowd Behavior for Computer Vision*. New York: Academic Press.

Nikhil N and Tran Morris B (2018) Convolutional neural network for trajectory prediction. In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Noe B and Collins N (2000) Variable structure interacting multiple-model filter (VS-IMM) for tracking targets with transportation network constraints. In: *SPIE Proceedings of Signal and Data Processing of Small Targets*, Vol. 4048.

Oh S, Hoogs A, Perera A, et al. (2011) A large-scale benchmark dataset for event recognition in surveillance video. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3153–3160.

Oli S, L'Esperance B and Gupta K (2013) Human motion behaviour aware planner (HMBAP) for path planning in dynamic human environments. In: *Proceedings of the IEEE International Conference on Advanced Robotics (ICAR)*, pp. 1–7.

Osa T, Pajarinen J, Neumann G, Bagnell JA, Abbeel P and Peters J (2018) An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics* 7(1–2): 1–179.

Palmieri L, Kucner TP, Magnusson M, Lilienthal AJ and Arras KO (2017) Kinodynamic motion planning on Gaussian mixture fields. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 6176–6181.

Pan X, He Y, Wang H, Xiong W and Peng X (2016) Mining regular behaviors based on multidimensional trajectories. *Expert Systems with Applications* 66: 106–113.

Pannetier B, Benameur K, Nimier V and Rombaut M (2005) VS-IMM using road map information for a ground target tracking. In: *Proceedings of the IEEE International Conference on Information Fusion (Fusion)*.

Paris S, Pettré J and Donikian S (2007) Pedestrian reactive navigation for crowd simulation: a predictive approach. *Computer Graphics Forum* 26: 665–674.

Park SH, Kim B, Kang CM, Chung CC and Choi JW (2018) Sequence-to-sequence prediction of vehicle trajectory via LSTM encoder-decoder architecture. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*. IEEE, pp. 1672–1678.

Pei Z, Qi X, Zhang Y, Ma M and Yang YH (2019) Human trajectory prediction in crowded scene using social-affinity long short-term memory. *Pattern Recognition* 93: 273–282.

Pellegrini S, Ess A, Schindler K and van Gool L (2009) You'll never walk alone: Modeling social behavior for multi-target tracking. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 261–268.

Pellegrini S, Ess A and van Gool L (2010) Improving data association by joint modeling of pedestrian trajectories and groupings. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 452–465.

Pentland A and Liu A (1999) Modeling and prediction of human behavior. *Neural Computation* 11(1): 229–242.

Petrich D, Dang T, Kasper D, Breuel G and Stiller C (2013) Map-based long term motion prediction for vehicles in traffic environments. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 2166–2172.

Pettré J, Ondřej J, Olivier AH, Cretual A and Donikian S (2009) Experiment-based modeling, simulation and validation of interactions between virtual walkers. In: *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 189–198.

Pfeiffer M, Paolo G, Sommer H, Nieto J, Siegwart R and Cadena C (2018) A data-driven model for interaction-aware pedestrian motion prediction in object cluttered environments. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8.

Pfeiffer M, Schwesinger U, Sommer H, Galceran E and Siegwart R (2016) Predicting actions to act predictably: Cooperative partial motion planning with maximum entropy models. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 2096–2101.

Piciarelli C, Foresti GL and Snidaro L (2005) Trajectory clustering and its applications for video surveillance. In: *IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, pp. 40–45.

Poiesi F and Cavallaro A (2015) Predicting and recognizing human interactions in public spaces. *Journal of Real-Time Image Processing* 10(4): 785–803.

Pool EA, Kooij JF and Gavrila DM (2019) Context-based cyclist path prediction using recurrent neural networks. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*. IEEE, pp. 824–830.

Pool EAI, Kooij JFP and Gavrila DM (2017) Using road topology to improve cyclist path prediction. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 289–296.

Previtali F, Bordallo A, Iocchi L and Ramamoorthy S (2016) Predicting future agent motions for dynamic environments. In: *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 94–99.

Qiu F and Hu X (2010) Modeling group structures in pedestrian crowd simulation. *Simulation Modelling Practice and Theory* 18(2): 190–205.

Quehl J, Hu H, Taş Öc, Rehder E and Lauer M (2017) How good is my prediction? Finding a similarity measure for trajectory prediction evaluation. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 1–6.

Quintero R, Almeida J, Llorca DF and Sotelo MA (2014) Pedestrian path prediction using body language traits. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 317–323.

Radwan N, Valada A and Burgard W (2018) Multimodal interaction-aware motion prediction for autonomous street crossing. *arXiv preprint arXiv:1808.06887*.

Raipuria G, Gaisser F and Jonker PP (2018) Road infrastructure indicators for trajectory prediction. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 537–543.

Rasouli A and Tsotsos JK (2019) Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Transactions on Intelligent Transport Systems*, in press.

Rehder E and Klöden H (2015) Goal-directed pedestrian prediction. In: *Proceedings of the International Conference on Computer Vision Workshop*, pp. 139–147.

Rehder E, Wirth F, Lauer M and Stiller C (2018) Pedestrian prediction by planning using deep neural networks. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–5.

Rhinehart N, Kitani K and Vernaza P (2018a) R2P2: A ReparameteRized Pushforward Policy for diverse, precise generative path forecasting. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 772–788.

Rhinehart N, McAllister R, Kitani K and Levine S (2019) PRECOG: Prediction conditioned on goals in visual multi-agent settings. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Rhinehart N, McAllister R and Levine S (2018b) Deep imitative models for flexible inference, planning, and control. *arXiv preprint arXiv:1810.06544*.

Ridel D, Deo N, Wolf D and Trivedi M (2019) Scene compliant trajectory forecast with agent-centric spatio-temporal grids. *arXiv preprint arXiv:1909.07507*.

Ridel D, Rehder E, Lauer M, Stiller C and Wolf D (2018) A literature review on the prediction of pedestrian behavior in urban scenarios. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*. IEEE, pp. 3105–3112.

Robicquet A, Sadeghian A, Alahi A and Savarese S (2016) Learning social etiquette: Human trajectory understanding in crowded scenes. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 549–565.

Rösmann C, Hoffmann F and Bertram T (2015) Timed-elastic-bands for time-optimal point-to-point nonlinear model predictive control. In: *Proceedings of the European Control Conference (ECC)*. IEEE, pp. 3352–3357.

Rösmann C, Oeljeklaus M, Hoffmann F and Bertram T (2017) Online trajectory prediction and planning for social robot navigation. In: *Proceedings of the IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, pp. 1255–1260.

Roth M, Flohr F and Gavrila DM (2016) Driver and pedestrian awareness-based collision risk analysis. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 454–459.

Rudenko A, Kucner TP, Swaminathan CS, Chadalavada RT, Arras KO and Lilienthal AJ (2019) THÖR: Human–robot navigation data collection and accurate motion trajectories dataset. *arXiv preprint arXiv:1909.04403*.

Rudenko A, Palmieri L and Arras KO (2017) Predictive planning for a mobile robot in human environments. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Workshop on AI Planning and Robotics*.

Rudenko A, Palmieri L and Arras KO (2018a) Joint prediction of human motion using a planning-based social force approach. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–7.

Rudenko A, Palmieri L, Lilienthal AJ and Arras KO (2018b) Human motion prediction under social grouping constraints. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*.

Russell SJ and Norvig P (2016) *Artificial Intelligence: A Modern Approach*. New York: Pearson.

Sadeghian A, Kosaraju V, Gupta A, Savarese S and Alahi A (2018) TrajNet: Towards a benchmark for human trajectory prediction. *arXiv preprint*.

Sadeghian A, Kosaraju V, Sadeghian A, Hirose N and Savarese S (2019) SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1349–1358.

Saleh K, Hossny M and Nahavandi S (2018a) Cyclist trajectory prediction using bidirectional recurrent neural networks. In: *Australasian Joint Conference on Artificial Intelligence*. Berlin: Springer, pp. 284–295.

Saleh K, Hossny M and Nahavandi S (2018b) Intent prediction of pedestrians via motion trajectories using stacked recurrent neural networks. *IEEE Transactions on Intelligent Vehicles* 3(4): 414–424.

Saleh K, Hossny M and Nahavandi S (2019) Contextual recurrent predictive model for long-term intent prediction of vulnerable road users. *IEEE Transactions on Intelligent Transport Systems (TITS)*.

Schmerling E, Leung K, Vollprecht W and Pavone M (2018) Multimodal probabilistic model-based planning for human–robot interaction. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9.

Schneemann F and Heinemann P (2016) Context-based detection of pedestrian crossing intention for autonomous driving in urban environments. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 2243–2248.

Schneider N and Gavrila DM (2013) Pedestrian path prediction with recursive Bayesian filters: A comparative study. In: *Proceedings of the German Conference on Pattern Recognition*. Berlin: Springer, pp. 174–183.

Schöller C, Aravantinos V, Lay F and Knoll A (2019) The simpler the better: Constant velocity for pedestrian motion prediction. *arXiv preprint arXiv:1903.07933*.

Schrijver A (2012) On the history of the shortest path problem. In: *Documenta Mathematica*, extra volume, pp. 155–167.

Schubert R, Richter E and Wanielik G (2008) Comparison and evaluation of advanced motion models for vehicle tracking. In: *Proceedings of the IEEE International Conference on Information Fusion (Fusion)*, pp. 1–6.

Schulz AT and Stiefelhagen R (2015) A controlled interactive multiple model filter for combined pedestrian intention recognition and path prediction. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*, pp. 173–178.

Seitz M, Köster G and Pfaffinger A (2012) Pedestrian group behavior in a cellular automaton. In: *Pedestrian and Evacuation Dynamics*. Berlin: Springer, pp. 807–814.

Sethian JA (1996) A fast marching level set method for monotonically advancing fronts. *Proceedings of the National Academy of Sciences* 93(4): 1591–1595.

Shalev-Shwartz S, Ben-Zrihem N, Cohen A and Shashua A (2016) Long-term planning by short-term prediction. *arXiv preprint arXiv:1602.01580*.

Shea PJ, Zadra T, Klamer DM, Frangione E and Brouillard R (2000) Improved state estimation through use of roads in ground tracking. In: *SPIE Proceedings of Signal and Data Processing of Small Targets*, Vol. 4048, pp. 321–333.

Shen M, Habibi G and How JP (2018) Transferable pedestrian motion prediction models at intersections. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 4547–4553.

Shi X, Shao X, Guo Z, Wu G, Zhang H and Shibasaki R (2019) Pedestrian trajectory prediction in extremely crowded scenarios. *Sensors* 19(5): 1223.

Simon D (2010) Kalman filtering with state constraints: A survey of linear and nonlinear algorithms. *IET Control Theory and Applications* 4: 1303–1318.

Singh H, Arter R, Dodd L, Langston P, Lester E and Drury J (2009) Modelling subgroup behaviour in crowd dynamics DEM simulation. *Applied Mathematical Modelling* 33(12): 4408–4423.

Srikanth S, Ansari JA, Ram KR, Sharma S, Murthy JK and Krishna KM (2019) INFER: INtermediate representations for FuturE pRediction. *arXiv preprint arXiv:1903.10641*.

Su H, Zhu J, Dong Y and Zhang B (2017) Forecast the plausible paths in crowd scenes. In: *Proceedings of IJCAI*, Vol. 1, p. 2.

Sumpter N and Bulpitt A (2000) Learning spatio-temporal patterns for predicting object behaviour. *Image and Vision Computing* 18(9): 697–704.

Sun L, Yan Z, Mellado SM, Hanheide M and Duckett T (2018) 3DOF pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–7.

Sung C, Feldman D and Rus D (2012) Trajectory clustering for motion prediction. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 1547–1552.

Suraj MS, Grimmett H, Platinskỳ L and Ondrúška P (2018) Predicting trajectories of vehicles using large-scale motion priors.

In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 1639–1644.

Tadokoro S, Ishikawa Y, Takebe T and Takamori T (1993) Stochastic prediction of human motion and control of robots in the service of human. In: *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics (SMC)*, Vol. 1, pp. 503–508.

Tang C and Salakhutdinov RR (2019) Multiple futures prediction. In: *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15398–15408.

Tay MKC and Laugier C (2008) Modelling smooth paths using gaussian processes. In: *Results of the International Conference on Field and Service Robotics*. Berlin: Springer, pp. 381–390.

Thiede LA and Brahma PP (2019) Analyzing the variety loss in the context of probabilistic trajectory prediction. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 9954–9963.

Thompson S, Horiuchi T and Kagami S (2009) A probabilistic model of human motion and navigation intent for mobile robot path planning. In: *Proceedings of the IEEE International Conference on Autonomous Robots and Agents (ICARA)*, pp. 663–668.

Tran Q and Firl J (2014) Online maneuver recognition and multimodal trajectory prediction for intersection assistance using non-parametric regression. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 918–923.

Trautman P and Krause A (2010) Unfreezing the robot: Navigation in dense, interacting crowds. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 797–803.

Trautman P, Ma J, Murray RM and Krause A (2013) Robot navigation in dense human crowds: The case for cooperation. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2153–2160.

Treiber M, Hennecke A and Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Physical Review E* 62(2): 1805.

Unhelkar VV, Pérez-D'Arpino C, Stirling L and Shah JA (2015) Human–robot co-navigation using anticipatory indicators of human walking motion. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6183–6190.

van den Berg J, Lin M and Manocha D (2008) Reciprocal velocity obstacles for real-time multi-agent navigation. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1928–1935.

van Den Berg J, Patil S, Sewall J, Manocha D and Lin M (2008) Interactive navigation of multiple agents in crowded environments. In: *Proceedings of the ACM Symposium on Interactive 3D Graphics and Games*, pp. 139–147.

van der Heiden T, Nagaraja NS, Weiss C and Gavves E (2019) SafeCritic: Collision-aware trajectory prediction. *arXiv preprint arXiv:1910.06673*.

Varshneya D and Srinivasaraghavan G (2017) Human trajectory prediction using spatially aware deep attention models. *arXiv preprint arXiv:1705.09436*.

Vasishta P, Vaufreydaz D and Spalanzani A (2017) Natural vision based method for predicting pedestrian behaviour in urban environments. In: *Proceedings of the IEEE International Conference on Intelligent Transport Systems (ITSC)*.

Vasishta P, Vaufreydaz D and Spalanzani A (2018) Building prior knowledge: A Markov based pedestrian prediction model using urban environmental data. In: *Proceedings of the International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1–12.

Vasquez D (2016) Novel planning-based algorithms for human motion prediction. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3317–3322.

Vasquez D, Fraichard T, Aycard O and Laugier C (2008) Intentional motion on-line learning and prediction. *Machine Vision and Applications* 19(5): 411–425.

Vasquez D, Fraichard T and Laugier C (2009) Incremental learning of statistical motion patterns with growing hidden Markov models. *IEEE Transactions on Intelligent Transport Systems* 10(3): 403–416.

Vemula A, Muelling K and Oh J (2017) Modeling cooperative navigation in dense human crowds. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1685–1692.

Vemula A, Muelling K and Oh J (2018) Social Attention: Modeling attention in human crowds. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Vintr T, Molina S, Senanayake R, et al. (2019) Time-varying pedestrian flow models for service robots. In: *Proceedings of the European Conference on Mobile Robots (ECMR)*. IEEE, pp. 1–7.

Völz B, Mielenz H, Siegwart R and Nieto J (2016) Predicting pedestrian crossing using quantile regression forests. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 426–432.

Walker J, Doersch C, Gupta A and Hebert M (2016) An uncertain future: Forecasting from static images using variational autoencoders. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Berlin: Springer, pp. 835–851.

Walker J, Gupta A and Hebert M (2014) Patch to the future: Unsupervised visual prediction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3302–3309.

Wang Z, Jensfelt P and Folkesson J (2015) Modeling spatial-temporal dynamics of human movements for predicting future trajectories. In: *Workshop Proceedings of the AAAI Conference on Artificial Intelligence "Knowledge, Skill, and Behavior Transfer in Autonomous Robots"*.

Wang Z, Jensfelt P and Folkesson J (2016) Building a human behavior map from local observations. In: *Proceedings of the IEEE International Symposium on Robot and Human Interactive Commication (RO-MAN)*, pp. 64–70.

Wu J, Ruenz J and Althoff M (2018) Probabilistic map-based pedestrian motion prediction taking traffic participants into consideration. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 1285–1292.

Xiao S, Wang Z and Folkesson J (2015) Unsupervised robot learning to predict person motion. In: *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 691–696.

Xie D, Todorovic S and Zhu SC (2013) Inferring "dark matter" and "dark energy" from videos. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2224–2231.

Xie G, Gao H, Qian L, Huang B, Li K and Wang J (2018) Vehicle trajectory prediction by integrating physics-and maneuver-based approaches using interactive multiple models. *IEEE Transactions on Industrial Electronics* 65(7): 5999–6008.

Xu Y, Piao Z and Gao S (2018) Encoding crowd interaction with deep neural network for pedestrian trajectory prediction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5275–5284.

Xue H, Huynh D and Reynolds M (2019) Location-velocity attention for pedestrian trajectory prediction. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 2038–2047.

Xue H, Huynh DQ and Reynolds M (2017) Bi-prediction: pedestrian trajectory prediction based on bidirectional LSTM classification. In: *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, pp. 1–8.

Xue H, Huynh DQ and Reynolds M (2018) SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction. In: *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 1186–1194.

Yamaguchi K, Berg AC, Ortiz LE and Berg TL (2011) Who are you with and where are you going? In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1345–1352.

Yan X, Kakadiaris IA and Shah SK (2014) Modeling local behavior for predicting social interactions towards human tracking. *Pattern Recognition* 47(4): 1626–1641.

Yan Z, Duckett T and Bellotto N (2017) Online learning for human classification in 3D LiDAR-based tracking. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 864–871.

Yang C, Bakich M and Blasch E (2005) Nonlinear constrained tracking of targets on roads. In: *Proceedings of the IEEE International Conference on Information Fusion (Fusion)*.

Yang C and Blasch E (2008) Fusion of tracks with road constraints. *Journal of Advances in Information Fusion* 3: 1.

Yen HC, Huang HP and Chung SY (2008) Goal-directed pedestrian model for long-term motion prediction with application to robot motion planning. In: *Proceedings of the IEEE Workshop on Advanced Robotics and Its Social Impacts*, pp. 1–6.

Yi S, Li H and Wang X (2016) Pedestrian behavior modeling from stationary crowds with applications to intelligent surveillance. *IEEE Transactions on Image Processing* 25(9): 4354–4368.

Yoo Y, Yun K, Yun S, Hong J, Jeong H and Young Choi J (2016) Visual path prediction in complex scenes with crowded moving objects. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2668–2677.

Zanlungo F, Ikeda T and Kanda T (2011) Social force model with explicit collision prediction. *EPL (Europhysics Letters)* 93(6): 68005.

Zechel P, Streiter R, Bogenberger K and Göhner U (2019) Pedestrian occupancy prediction for autonomous vehicles. In: *IEEE International Conference on Robotic Computing (IRC)*. IEEE, pp. 230–235.

Zernetsch S, Kohnen S, Goldhammer M, Doll K and Sick B (2016) Trajectory prediction of cyclists using a physical model and an artificial neural network. In: *Proceedings of the IEEE Intelligent Vehicle Symposium (IV)*, pp. 833–838.

Zhan E, Zheng S, Yue Y and Lucey P (2018) Generative multi-agent behavioral cloning. *arXiv preprint arXiv:1803.07612*.

Zhang P, Ouyang W, Zhang P, Xue J and Zheng N (2019) SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12085–12094.

Zhang Z, Huang K and Tan T (2006) Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In: *18th International Conference on Pattern Recognition, 2006 (ICPR 2006)*, Vol. 3. IEEE, pp. 1135–1138.

Zhao T, Xu Y, Monfort M, et al. (2019) Multi-agent tensor fusion for contextual trajectory prediction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12126–12134.

Zheng S, Yue Y and Hobbs J (2016) Generating long-term trajectories using deep hierarchical networks. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 1543–1551.

Zheng Y (2015) Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology* 6(3): 29.

Zhou B, Tang X and Wang X (2015) Learning collective crowd behaviors with dynamic pedestrian-agents. *International Journal of Computer Vision* 111(1): 50–68.

Zhou B, Wang X and Tang X (2012) Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2871–2878.

Zhu Q (1991) Hidden Markov Model for dynamic obstacle avoidance of mobile robot navigation. *IEEE Transactions on Robotics and Automation* 7(3): 390–397.

Ziebart BD, Ratliff N, Gallagher G, et al. (2009) Planning-based prediction for pedestrians. In: *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 3931–3936.