

# Learning and Inferring “Dark Matter” and Predicting Human Intents and Trajectories in Videos

Dan Xie, Tianmin Shu, Sinisa Todorovic and Song-Chun Zhu

**Abstract**—This paper presents a method for localizing functional objects and predicting human intents and trajectories in surveillance videos of public spaces, under no supervision in training. People in public spaces are expected to intentionally take shortest paths (subject to obstacles) toward certain objects (e.g. vending machine, picnic table, dumpster etc.) where they can satisfy certain needs (e.g., quench thirst). Since these objects are typically very small or heavily occluded, they cannot be inferred by their visual appearance but indirectly by their influence on people’s trajectories. Therefore, we call them “dark matter”, by analogy to cosmology, since their presence can only be observed as attractive or repulsive “fields” in the public space. A person in the scene is modeled as an intelligent agent engaged in one of the “fields” selected depending his/her intent. An agent’s trajectory is derived from an Agent-based Lagrangian Mechanics. The agents can change their intents in the middle of motion and thus alter the trajectory. For evaluation, we compiled and annotated a new dataset. The results demonstrate our effectiveness in predicting human intent behaviors and trajectories, and localizing and discovering distinct types of “dark matter” in wide public spaces.

**Index Terms**—scene understanding, video analysis, functional objects, intents modeling, trajectory projection



## 1 INTRODUCTION

### 1.1 Motivation and Objective

This paper addresses inference of why and how people move in surveillance videos of public spaces (e.g., park, campus), under no supervision in training. Regarding the “why”, we expect that people typically have certain needs (e.g., to quench thirst, satiate hunger, get some rest), and hence intentionally move toward certain destinations in the scene where these needs can be satisfied (e.g., vending machine, food truck, bench). Regarding the “how”, we make the assumption that people take shortest paths to intended destinations, while avoiding obstacles and non-walkable surfaces. We also consider three types of human behavior, including: “single intent” when a person reaches the destination and stops, “sequential intent” when a person sequentially visits several functional objects (e.g., buy food at the food-truck, and go to a bench to have lunch), and “change of intent” when a person initially heads to one goal but then changes the goal (e.g. because the line in front of the food-truck is too long).

The answers to the above “why” and “how” are important, since they can be used toward a “deeper” scene and event understanding than that considered by related work, in terms of predicting human trajectories in the future, reasoning about latent human intents and behavior, and localizing functional objects and non-walkable areas in the scene.

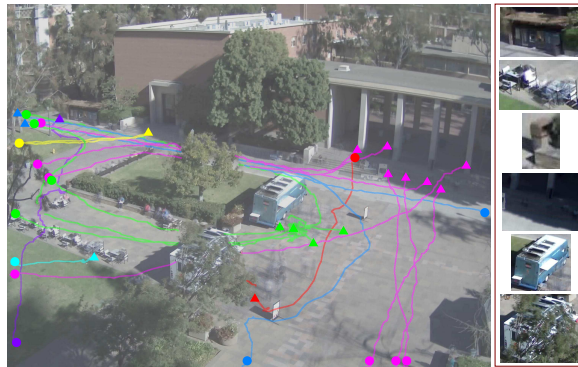


Fig. 1: (*left*) People’s trajectories are color-coded by their shared goal destination. The triangles denote destinations, and the dots denote start positions of the trajectories. E.g., people may be heading toward the food-truck to buy food (green), or the vending machine to quench thirst (blue). (*right*) Due to low resolution, poor lighting, and occlusions, objects at the destinations are very difficult to detect only based on their appearance and shape.

It is worth noting that destinations of human trajectories are typically occupied by objects that are poorly visible even by a human eye, due to the low-resolution of our surveillance videos, as illustrated in Fig. 1. We call these objects “dark matter”, because they are distinguishable from other objects primarily by the functionality to attract or repel people, not by their appearance. A detection of such objects based on appearance would be unreliable. We use this terminology to draw an analogy to cosmology, where existence and properties of dark matter are hypothesized and inferred from its gravitational effects on visible matter. Analogously, we consider poorly visible objects at destinations of human trajectories as different types of “dark

- D. Xie and T. Shu are with Department of Statistics, University of California, Los Angeles. Email: {xiedan,tianmin.shu}@ucla.edu.
- S. Todorovic is with School of EECS, Oregon State University. Email: sinisa@oregonstate.edu.
- S.-C. Zhu is with Department of Statistics and Computer Science, University of California, Los Angeles. Email: sczhu@stat.ucla.edu.

Examples of “dark matter”	Human need
Vending machine / Food truck / Table	Hunger
Water fountain / Vending machine	Thirst
ATM / Bank	Money
Chair / Table / Bench / Grass	Rest
News stand / Ad billboard	Information
Trash can	Hygiene
Bush / Tree	Shade from the sun

TABLE 1: Examples of human needs and objects that can satisfy these needs in the context of a public space. These objects appear as “dark matter” attracting people to approach them, or repelling people to stay away from them.

matter” exerting attraction and repulsion forces on people. Each type is defined probabilistically by the corresponding human-trajectory pattern around the “dark matter”. Tab. 1 lists examples of human needs and objects with “dark-matter” functionality considered in this paper.

**Problem statement:** Given a video of a public space, our problem involves *unsupervised* prediction of:

- Human intents, trajectories, and behaviors,
- Locations of “dark matter” and non-walkable surfaces, i.e., functional map of the scene,
- Attraction or repulsion force fields of the localized “dark matter”.

In this paper, we also consider unsupervised discovery of different types of “dark matter” in a given set of surveillance videos. As our experiments demonstrate, each discovered type groups a certain semantically meaningful class of objects with the corresponding function in the scene (e.g., stairs and entrance doors of buildings form a type of “dark matter” where people exit the scene).

This work focuses on the unsupervised setting where ground truth for objects in the scene representing “dark matter” and their functionality is not available in training. Studying such a setting is important, since providing ground truth about functionality of objects would be very difficult in our video domain, in part, due to the low video resolution and top-down views. Another difficulty for ground-truth annotation is that functionality of objects is not tightly correlated with their semantic classes, because instances of the same object may have different functionality in our scenes (e.g., a bench may attract people to get some rest, or repel them if freshly painted).

**Key contribution** of this paper involves a joint representation and inference of:

- Visible domain— traditional recognition categories: objects, scenes, actions and events; and
- Functional domain — higher level cognition concepts: fluent, causality, intents, attractions and physics.

To formulate this problem, we leverage the framework of Lagrange mechanics, and introduce the concept of *field*, analogous to gravitational field in physics. Each “dark matter” and non-walkable surface in the scene generates an attraction (positive) and repulsion (negative) field. Thus, we view the scene as a physical system populated by particle-agents who move in many layers of “dark-energy” fields.

Unlike inanimate particles, each agent can intentionally select a particular force field to affect its motions, and thus define the minimum-energy Dijkstra path toward the corresponding source “dark matter”. In the following, we introduce the main steps of our approach.

## 1.2 Overview of Our Approach

Fig. 2 illustrates main steps of our approach.

**Tracking.** Given a video, we first extract people’s trajectories using the state-of-the-art multitarget tracker of [1] and the low-level 3D scene reconstruction of [2]. While the tracker and 3D scene reconstruction perform well, they may yield noisy results. Also, these results represent only partial observations, since the tracks of most people in the given video are not fully observable, but get cut out at the end. These noisy, partial observations are used as input features to our inference.

**Bayesian framework.** Uncertainty is handled by specifying a joint pdf of observations, latent layout of non-walkable surfaces and functional objects, and people’s intents and trajectories. Our model is based on the following assumptions. People are expected to have only one goal destination at a time, and be familiar with the scene layout (e.g., from previous experience), such that they can optimize their trajectory as a shortest path toward the intended functional object, subject to the constraint map of non-walkable surfaces. We consider three types of intent behavior. A person may change the intent and decide to switch to another goal destination, have only a single intent, or want to sequentially reach several functional objects.

**Agent-based Lagrangian Mechanics.** Our Bayesian framework leverages the Lagrangian mechanics (LM) by treating the scene as a physics system where people can be viewed as charged particles moving along the mixture of repulsion and attraction energy fields generated by obstacles and functional objects. The classical LM, however, is not directly applicable to our domain, because it deterministically applies the principle of Least Action, and thus provides a poor model of human behavior.

We extend LM to an agent-based Lagrangian mechanics (ALM) which accounts for latent human intentions. Specifically, in ALM, people can be viewed as charged particle-agents with capability to intentionally select one of the latent fields, which in turn guides their motions by the principle of Least Action.

**Inference.** We use the data-driven Markov Chain Monte Carlo (MCMC) for inference [3], [4]. In each iteration, the MCMC probabilistically samples the number and locations of obstacles and sources of “dark energy”, and people’s intents. This, in turn, uniquely identifies the “dark energy” fields in the scene. Each person’s trajectory is estimated as the globally optimal Dijkstra path in these fields, subject to obstacle constraints. The predicted trajectories are used to estimate if they arose from “single”, “sequential” or “change” of human intents. In this paper, we consider two inference settings: *offline* and *online*. The former first infers the layout of “dark matter” and obstacles in the scene as

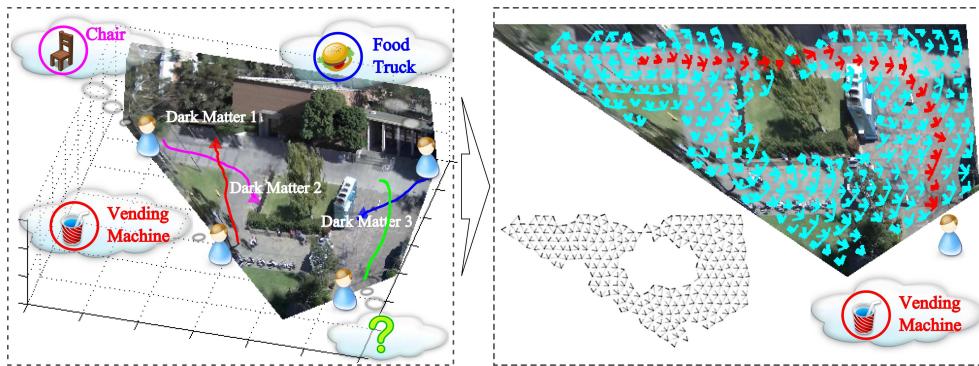


Fig. 2: An example video where people driven by latent needs move toward functional objects where these needs can be satisfied (i.e., “dark matter”). (Right) A zoomed-in top-down view of the scene and our actual results of: (a) Inferring and localizing the person’s goal destination; (b) Predicting the person’s full trajectory (red); (c) Estimating the force field affecting the person (the blue arrows, where their thickness indicates the force magnitude; the black arrows represent another visualization of the same field.); and (d) Estimating the constraint map of non-walkable areas and obstacles in the scene (the “holes” in the field of blue arrows and the field of black arrows).

well as people’s intents, and then fixes these estimates for predicting Dijkstra trajectories and human intent behavior. The latter sequentially estimates both people’s intents and human trajectories frame by frame, where the estimation for frame  $t$  uses all previous predictions.

We present experimental evaluation on challenging, real-world videos from the VIRAT [5], UCLA Courtyard [6], UCLA Aerial Event [7] datasets, as well as on our five new videos of public squares. Our ground truth annotations and the new dataset will be made public. The results demonstrate our effectiveness in predicting human intent behaviors and trajectories, and localizing functional objects, as well as discovering distinct functional classes of objects by clustering human motion behavior in the vicinity of functional objects. Since localizing functional objects in videos is a new problem, we compare with existing approaches only in terms of predicting human trajectories. The results show that we outperform prior work on VIRAT and UCLA Courtyard datasets.

### 1.3 Relationship to Prior Work

This section reviews three related research streams in the literature, including the work on functionality recognition, human tracking, and prediction of events. For each stream, we also point out our differences and contributions.

**Functionality recognition.** Recent work has demonstrated that performance in object and human activity recognition can be improved by reasoning about functionality of objects. Functionality is typically defined as an object’s capability to satisfy certain human needs, which in turn triggers corresponding human behavior. E.g., reasoning about how people handle and manipulate small objects can improve accuracy of recognizing calculators or cell-phones [8], [9]. Some other object classes can be directly recognized by estimating how people interact with the objects [10], rather than using common appearance features. This interaction can be between a person’s hands and the object [11], or between a human skeleton and the objects

[12]. Another example is the approach that successfully recognizes chairs among candidate objects observed in the image by using human-body poses as context for identifying whether the candidates have functionality “sittable” [13]. Similarly, video analysis can be improved by detecting and localizing functional scene elements, such as parking spaces, based on low-level appearance and local motion features [14]. The functionality of moving objects [15] and urban road environments [16] has been considered for advancing activity recognition.

As in the above approaches, we also resort to reasoning about functionality of objects based on human behavior and interactions with the objects, rather than use standard appearance-based features. Our key difference is that we *explicitly* model latent human intents which can modify an object’s functionality – specifically, in our domain, an object may simultaneously attract some people *and* repel others, depending on their intents.

**Human tracking and planning.** A survey of vision-based trajectory learning and analysis for surveillance is presented in [17]. The related approaches differ from ours in the following aspects. Estimations of: (a) Representative human motion patterns in (years’) long video footage [18], (b) Lagrangian particle dynamics of crowd flows [19], and (c) Optical-flow based dynamics of crowd behaviors [20] do not account for individual human intents. Reconstruction of an unobserved trajectory segment has been addressed only as finding the shortest path between the observed start and end points [21]. Early work also estimated a numeric potential field for robot path planning [22], but did not account for the agents free will to choose and change goal destinations along their paths. Optimal path search [23], and reinforcement learning and inverse reinforcement learning [24], [25], [26] was used for explicitly reasoning about people’s goals for predicting human trajectories. However, these approaches considered: i) Relatively sanitized settings with scenes that did not have many and large obstacles (e.g., parking lots); and ii) Limited set

of locations for people’s goals (e.g., along the boundary of the video frames). People’s trajectories have also been estimated based on inferring social interactions [27], [28], [29], [30], or detecting objects in egocentric videos [31]. However, these approaches critically depend on domain knowledge. For example, the approaches of [26] and [31] use appearance-based object detectors, learned on training data, for predicting trajectories. In contrast, we are not in a position to apply appearance-based object detectors for identifying hidden functional objects, due to the low resolution of our videos. Finally, a Mixture of Kalman Filters has been used to cluster *smooth* human trajectories based on their dynamics and start and end points [32]. Instead of linear dynamics, we use the principle of Least Action, and formulate a globally optimal planning of the trajectories. This allows us to handle sudden turns and detours caused by obstacles or change of intent. Our novel formulation advances Lagrangian Mechanics.

Related to ours is prior work in cognitive science [33] aimed at inferring human goal destinations based on inverting a probabilistic generative model of goal-dependent plans from an incomplete sequence of human behavior. Also, similar to our MCMC sampling, the Wang-Landau Monte Carlo (WLMC) sampling is used in [4] for people tracking in order to handle abrupt motions.

**Prediction and early decision.** There is growing interest in action prediction [34], [35], [36], and early recognition of a single human activity [37] or a single structured event [38], [39]. These approaches are not aimed at predicting human trajectories, and are not suitable for our domain in which multiple activities may happen simultaneously. Also, some of them make the assumption that human activities are structured [38], [39] which is relatively rare in our surveillance videos of public spaces where people mostly just walk or remain still. Another difference is that we distinguish activities by human intent, rather than their semantic meaning. Some early recognition approaches do predict human trajectories [40], but use a deterministic vector field of people’s movements, whereas our “dark energy” fields are stochastic. In [41], an anticipatory temporal conditional random field (ATCRF) is used for predicting human activities based on object affordances. These activities are, however, defined at the human-body scale, and thus the approach cannot be easily applied to our wide-scene views. A linear dynamic system of [42], [32] models smooth trajectories of pedestrians in crowded scenes, and thus cannot handle sudden turns and detours caused by obstacles, as required in our setting. In graphics, relatively simplistic models of agents are used to simulate people’s trajectories in a virtual crowd [43], [44], [45], but cannot be easily extended to our surveillance domain. Unlike the above related work, we do not exploit appearance-based object detectors for localizing objects that can serve as possible people’s destinations in the scene.

**Extensions from our preliminary work.** We extend our preliminary work [46] by additionally: 1) Modeling and inferring “sequential” human intents and “change of intent” along the course of people’s trajectories; 2) Online

prediction of human intents and trajectories; 3) Clustering functional objects; and 4) Presenting the corresponding new empirical results. Neither change of intent nor “sequential” intents were considered in [46].

## 1.4 Contributions

This paper makes the following three contributions.

- *Agent-based Lagrangian Mechanics (ALM)*. We leverage the Lagrangian mechanics (LM) for modeling human motion in an outdoor scene as a physical system. The LM is extended to account for human free will to choose goal destinations and change intent.
- *Force-dynamic functional map*. We present a novel approach to modeling and estimating the force-dynamic functional map of a scene in the surveillance video.
- *Human intents*. We explicitly model latent human intents, and allow a person to change intent.

## 2 AGENT-BASED LAGRANGIAN MECHANICS

At the scale of large scenes such as courtyard, people are considered as “particles” whose shapes and dimensions are neglected, and their motion dynamics modeled within the framework of Lagrangian mechanics (LM) [47]. LM studies the motion of a particle with mass,  $m$ , at positions  $\mathbf{x}(t) = (x(t), y(t))$  and velocity,  $\dot{\mathbf{x}}(t)$ , in time  $t$ , in a force field  $\vec{F}(\mathbf{x}(t))$  affecting the motion of the particle. Particle motion in generalized coordinates system is determined by the Lagrangian function,  $L(\mathbf{x}, \dot{\mathbf{x}}, t)$ , defined as the kinetic energy of the entire physical system,  $\frac{1}{2}m\dot{\mathbf{x}}(t)^2$ , minus its potential energy,  $-\int_{\mathbf{x}} \vec{F}(\mathbf{x}(t))d\vec{\mathbf{x}}(t)$ ,

$$L(\mathbf{x}, \dot{\mathbf{x}}, t) = \frac{1}{2}m\dot{\mathbf{x}}(t)^2 + \int_{\mathbf{x}} \vec{F}(\mathbf{x}(t))d\vec{\mathbf{x}}(t). \quad (1)$$

Action in such a physical system is defined as the time integral of the Lagrangian of trajectory  $\mathbf{x}$  from  $t_1$  to  $t_2$ :  $\int_{t_1}^{t_2} L(\mathbf{x}, \dot{\mathbf{x}}, t)dt$ .

LM postulates that a particle’s trajectory,  $\Gamma(t_1, t_2) = [\mathbf{x}(t_1), \dots, \mathbf{x}(t_2)]$ , is governed by the principle of Least Action in a generalized coordinate system:

$$\Gamma(t_1, t_2) = \arg \min_{\mathbf{x}} \int_{t_1}^{t_2} L(\mathbf{x}, \dot{\mathbf{x}}, t)dt. \quad (2)$$

The classical LM is not directly applicable to our domain, because it considers inanimate objects. We extend LM in two key aspects, and thus derive the Agent-based Lagrangian mechanics (ALM). In ALM, a physical system consists of a set of force sources. Our first extension enables the particles to become agents with free will to select a particular force source from the set which can drive their motion. Our second extension endows the agents with knowledge about the layout map of the physical system. Consequently, by the principle of Least Action, they can globally optimize their shortest paths toward the selected force source, subject to the known layout of obstacles. These two extensions can be formalized as follows.

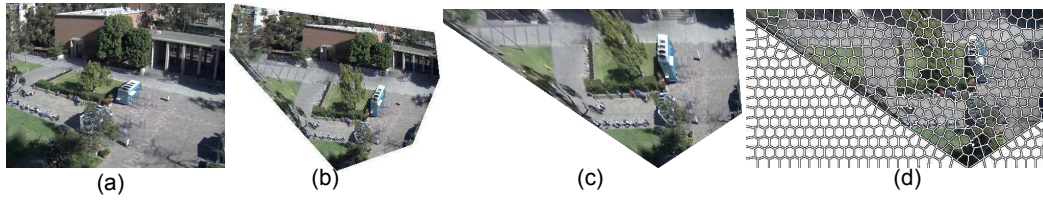


Fig. 3: (a) An example of a public space; (b) 3D reconstruction of the scene using the method of [2]; (c) Our estimation of the ground surface; and (d) Our inference is based on superpixels obtained using the method of [48].

Let  $i$ th agent choose  $j$ th source from the set of sources. Then,  $i$ 's action, i.e., trajectory is

$$\begin{aligned} & \Gamma_{ij}(t_1, t_2) \\ &= \arg \min_{\mathbf{x}} \int_{t_1}^{t_2} \left[ \frac{1}{2} m \dot{\mathbf{x}}(t)^2 + \int_{\mathbf{x}} \vec{F}_{ij}(\mathbf{x}(t)) d\vec{\mathbf{x}}(t) \right] dt, \quad (3) \\ & \text{s.t. } \mathbf{x}(t_1) = \mathbf{x}_i, \mathbf{x}(t_2) = \mathbf{x}_j. \end{aligned}$$

For solving the difficult optimization problem of (3) we resort to certain approximations, as explained below.

In our domain of public spaces, the agents cannot increase their speed without limit. Hence, every agent's speed is upper bounded by some maximum speed. Also, it seems reasonable to expect that accelerations or decelerations of people along their trajectories in a public space span negligibly short time intervals. Consequently, the first term in (3) is assumed to depend on a constant velocity of the agent, and thus does not affect estimation of  $\Gamma_{ij}(t_1, t_2)$ .

For simplicity, we allow the agent to make only discrete displacements over a lattice of scene locations  $\Lambda$  (e.g., representing centers of superpixels occupied by the ground surface in the scene), i.e.,  $d\vec{\mathbf{x}}(t) = \Delta\vec{\mathbf{x}}$ . Also, we expect that the agent is reasonable and always moves along the direction of  $\vec{F}_{ij}(\mathbf{x})$  at every location.

From (3) and above considerations, we derive:

$$\Gamma_{ij}(t_1, t_2) = \arg \min_{\Gamma \subset \Lambda} \sum_{\mathbf{x} \in \Gamma} |\vec{F}_{ij}(\mathbf{x}) \cdot \Delta\vec{\mathbf{x}}|, \quad (4)$$

such that  $\mathbf{x}(t_1) = \mathbf{x}_i$  and  $\mathbf{x}(t_2) = \mathbf{x}_j$ .

A globally optimal solution of (4) can be found with the Dijkstra algorithm. Note that the end location of the predicted  $\Gamma_{ij}(t_1, t_2)$  corresponds to the location of source  $j$ . It follows that estimating human trajectories can readily be used for estimating the functional map of the scene. To address uncertainty, this estimation is formulated within the Bayesian framework, as explained next.

### 3 PROBLEM FORMULATION

This section defines our probabilistic framework in terms of observable and latent variables. We first define all variables, and then specify their joint probability distribution. The notation is summarized in Tab. 2.

**Agents, Sources, Constraint Map:** The video shows agents,  $A = \{a_i : i = 1, \dots, M\}$ , and sources of “dark energy”,  $S = \{s_j : j = 1, \dots, N\}$ , occupying locations on the 2D lattice,  $\Lambda = \{\mathbf{x} = (x, y) : x, y \in \mathbb{Z}_+\}$ . Locations  $\mathbf{x} \in \Lambda$  may be walkable or non-walkable, as indicated by a constraint map,  $C = \{c(\mathbf{x}) : \forall \mathbf{x} \in \Lambda, c(\mathbf{x}) \in \{-1, 1\}\}$ ,

$\mathbf{x}$	A location $(x, y)$ on the ground plane
$\Gamma$	The trajectory of an agent
$a_i$	$i$ -th agent in the video
$A$	All agents in the video
$s_j$	Location of the $j$ -th source of “dark energy”
$S$	The locations of all sources of “dark energy”
$c(\mathbf{x})$	Indicator of walkability at $\mathbf{x}$ .
$C$	A constraint map
$\Lambda$	2D lattice
$\Lambda_1$	The set of walkable locations
$r_{ij}$	The relationship between agent $i$ and source $j$
$R$	The set of agent-goal relationships
$z_i$	Agent $i$ 's behavior type
$Z$	The set of the types of all agents's behavior
$W$	All latent variables
$\vec{F}^-(\mathbf{x})$	The repulsion force at location $\mathbf{x}$
$\vec{F}_j^+(\mathbf{x})$	The attraction force generated by source $s_j$ at $\mathbf{x}$

TABLE 2: Notation used in this paper.

where  $c(\mathbf{x}) = -1$ , if  $\mathbf{x}$  is non-walkable, and  $c(\mathbf{x}) = 1$ , otherwise. Walkable locations form the set  $\Lambda_1 = \{\mathbf{x} : \mathbf{x} \in \Lambda, c(\mathbf{x})=1\}$ .

**Intentions of Agents** are defined by the set of agent-goal relationships  $R = \{r_{ij}\}$ . When  $a_i$  wants to pursue  $s_j$ , we specify their relationship as  $r_{ij} = 1$ ; otherwise  $r_{ij} = 0$ . Note that  $a_i$  may pursue more than one source from  $S$  in a sequential manner, during the lifetime of the trajectory.

**Three Types of Agent Behavior:** In this paper, we consider three types of behavior: “single”, “sequential” and “change of intent”. We follow the definitions of intent types in [24]. The intent behavior of all agents is represented by a set of latent variables,  $Z = \{z_i\}$ . An agent  $a_i$  is assigned  $z_i = \text{“single”}$  when its intent is to achieve exactly one goal, and remain at the reached destination indefinitely. An agent  $a_i$  is assigned  $z_i = \text{“sequential”}$  when its intent is to achieve several goals along the trajectory. At each goal reached, the agent satisfies the corresponding need before moving to the next goal. In our videos, agents with sequential behavior typically visit no more than 3 destinations. An agent may also give up on the initial goal before reaching it, and switch to another goal. This defines  $z_i = \text{“change of intent”}$ . In our surveillance videos, we observe that “change of intent” happens relatively seldom, on average for only 1 agent in  $A$  in a given video. Also, we find that the goal-switching may occur equally likely at any moment, even when people are quite close to their initial goals (e.g., when seeing a long line in front of the food-truck).

**Repulsion Forces:** Sources  $S$  exert either repulsive or attractive forces on agents in  $A$ . Every non-walkable location  $\mathbf{x}' \in \Lambda \setminus \Lambda_1$  generates a repulsion force at agent's location  $\mathbf{x}$ ,  $\vec{F}_{\mathbf{x}'}^-(\mathbf{x})$ . The magnitude  $|\vec{F}_{\mathbf{x}'}^-(\mathbf{x})|$  is defined as the Mahalanobis distance in terms of the quadratic  $(\mathbf{x} - \mathbf{x}')^2$ , with covariance  $\Sigma = \sigma_r^2 \mathcal{I}$ , where  $\mathcal{I}$  is the

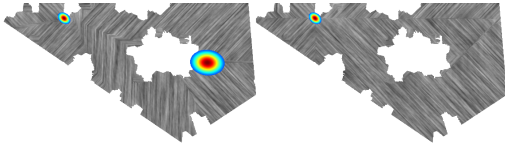


Fig. 4: Visualizations of the force field for the scene from Fig. 2. (left) In LM, particles are driven by a sum of all forces; the figure shows the resulting fields generated by only two sources. (right) In ALM, each agent selects a single force  $\vec{F}_j^+(\mathbf{x})$  to drive its motion; the figure shows that forces at all locations in the scene point toward the top left of the scene where the source is located. The white regions represent our estimates of obstacles. Repulsion forces are short ranged, with magnitudes too small to show here.

identity matrix, and  $\sigma_r^2 = 10^{-2}$  is empirically found as best. Thus, the magnitude  $|\vec{F}_{\mathbf{x}'}^-(\mathbf{x})|$  is large in the vicinity of non-walkable location  $\mathbf{x}'$ , but quickly falls to zero for locations farther away from  $\mathbf{x}'$ . This models our observation that a person may take a path that is very close to non-walkable areas, i.e., the repulsion force has a short-range effect on human trajectories. The sum of all repulsion forces arising from non-walkable areas in the scene gives the joint repulsion,  $\vec{F}^-(\mathbf{x}) = \sum_{\mathbf{x}' \in \Lambda \setminus \Lambda_1} \vec{F}_{\mathbf{x}'}^-(\mathbf{x})$ .

**Attraction Forces:** Each source  $\mathbf{s}_j \in S$  is capable of generating an attraction force,  $\vec{F}_j^+(\mathbf{x})$ , if selected as a goal destination by an agent at location  $\mathbf{x}$ . The magnitude  $|\vec{F}_j^+(\mathbf{x})|$  is specified as the Mahalanobis distance in terms of the quadratic  $(\mathbf{x} - \mathbf{x}_j)^2$ , with covariance  $\Sigma = \sigma_a^2 \mathcal{I}$  taken to be the same for all sources  $\mathbf{s}_j \in S$ , and  $\sigma_a^2 = 10^4$  is empirically found as best. This models our observation that people tend to first approach near-by functional objects, because, in part, reaching them requires less effort than approaching farther destinations. The attraction force is similar to the gravity force in physics whose magnitude becomes smaller as the distance increases.

**Net Force:** When  $a_i \in A$  selects  $\mathbf{s}_j \in S$ ,  $a_i$  is affected by the net force,  $\vec{F}_{ij}(\mathbf{x})$ , defined as:

$$\vec{F}_{ij}(\mathbf{x}) = \vec{F}^-(\mathbf{x}) + \vec{F}_j^+(\mathbf{x}). \quad (5)$$

From (5), we can more formally specify the difference between LM and our ALM, presented in Sec. 2. In LM, an agent would be affected by forces of all sources in  $S$ ,  $\vec{F}_{ij}^{\text{LM}}(\mathbf{x}) = \vec{F}^-(\mathbf{x}) + \sum_j \vec{F}_j^+(\mathbf{x})$ . In contrast, in ALM, an agent is affected by the force of a single selected source,  $\vec{F}_j^+(\mathbf{x})$ , along with the joint repulsion force. The difference between LM and our ALM is illustrated in Fig. 4.

**Trajectories of Agents:** In this paper, we make the assumption that we have access to noisy trajectories of agents, observed over a given time interval in the video,  $\Gamma' = \Gamma'(0, t_0) = \{\Gamma'_i(0, t_0) : i = 1, \dots, M\}$ . Given these observations, we define latent trajectories of agents for any time interval,  $(t_1, t_2)$ , including those in the future (i.e., unobserved intervals),  $\Gamma = \Gamma(t_1, t_2) = \{\Gamma_i(t_1, t_2) : i = 1, \dots, M\}$ . Each trajectory  $\Gamma_i$  is specified by accounting for one of the three possible behaviors of the agent as follows. Following the principle of Least Action Recall, as specified

in Sec. 2, an optimal trajectory  $\Gamma_{ij}(t_1, t_2) = [\mathbf{x}(t_1) = \mathbf{x}_i, \dots, \mathbf{x}(t_2) = \mathbf{x}_j]$  of  $a_i$  at location  $\mathbf{x}_i$  moving toward  $\mathbf{s}_j$  at location  $\mathbf{x}_j$  minimizes the energy  $\sum_{\mathbf{x} \in \Gamma_{ij}} |\vec{F}_{ij}(\mathbf{x}) \cdot \Delta \mathbf{x}|$ . Dropping notation for time, we extend this formulation to account for the agent’s behavior as

$$\Gamma_i = \sum_j \Gamma_{ij} = \arg \min_{\Gamma \subset \Lambda} \sum_j \sum_{\mathbf{x} \in \Gamma} |\vec{F}_{ij}(\mathbf{x}) \cdot \Delta \mathbf{x}|, \quad (6)$$

where the summation over  $j$  uses: (i) only one source for “single” intent (i.e.,  $\Gamma_i = \Gamma_{ij}$  when  $r_{ij} = 1$ ), (ii) two sources for “change of intent”, and (iii) maximally  $n$  sources for “sequential” behavior. Note that for the “sequential” behavior the minimization in (6) is constrained such that the trajectory must *sequentially* pass through locations  $\mathbf{x}_j$  of all sources  $\mathbf{s}_j$  pursued by the agent.

**The Probabilistic Model:** Using the aforementioned definitions of variables in our framework, we define the joint posterior distribution of latent variables  $W = \{C, S, R, Z, \Gamma\}$  given the observed trajectories of agents  $\Gamma' = \{\Gamma'_i\}$  and appearance features  $I$  in the video as

$$P(W|\Gamma', I) \propto P(C, S, R, Z)P(\Gamma, \Gamma'|C, S, R, Z)P(I|C), \quad (7)$$

where the joint prior is specified as

$$P(C, S, R, Z) = P(C)P(S|C)P(R|S)P(Z), \quad (8)$$

with distributions  $P(C)$ ,  $P(S|C)$ ,  $P(R|S)$ ,  $P(Z)$ , and  $P(I|C)$  defined in Sec. 4. For modeling the joint likelihood of trajectories in (7), we use the naïve Bayes model:

$$P(\Gamma, \Gamma'|C, S, R, Z) = \prod_{i=1}^M P(\Gamma_i|C, S, R, Z), \quad (9)$$

where  $P(\Gamma_i|C, S, R, Z)$  is also defined in Sec. 4.

In the following section, we define the priors and likelihoods of our probabilistic model characterized by (7).

## 4 DISTRIBUTIONS OF OUR MODEL

This section defines the priors and likelihoods of (8)–(9).

**Smoothness of Constraint Map:** The prior  $P(C)$  enforces spatial smoothness of the constraint map  $C$  using the Ising random field:

$$P(C) \propto \exp \left[ \beta \sum_{\mathbf{x} \in \Lambda, \mathbf{x}' \in \partial \mathbf{x} \cap \Lambda} c(\mathbf{x})c(\mathbf{x}') \right], \quad \beta > 0. \quad (10)$$

**Likelihood of Appearance Features:** We model walkable locations in the scene in terms of appearance features  $I$  extracted from the video. The likelihood of  $I$  is defined as

$$P(I|C) = \prod_{\mathbf{x} \in \Lambda} P(\phi(\mathbf{x})|c(\mathbf{x})=1), \quad (11)$$

where  $\phi(\mathbf{x})$  is a feature descriptor vector consisting of: i) the RGB color at the scene location  $\mathbf{x}$ , and ii) the binary indicator if  $\mathbf{x}$  belongs to the ground surface of the 3D reconstructed scene.  $P(\phi(\mathbf{x})|c(\mathbf{x})=1)$  is specified as a 2-component Gaussian mixture model. Note that

$P(\phi(\mathbf{x})|c(\mathbf{x})=1)$  is directly estimated on our given (single) video with latent  $c(\mathbf{x})$ , not using training data.

**Modeling Spatial Layout of Dark Matter:** The latent sources of “dark energy”,  $\mathbf{s}_j \in S$ , are characterized by location on the 2D lattice,  $\boldsymbol{\mu}_j \in \Lambda$ , and  $2 \times 2$  spatial covariance matrix  $\Sigma_j$  of  $\mathbf{s}_j$ ’s force field:

$$S = \{\mathbf{s}_j = (\boldsymbol{\mu}_j, \Sigma_j) : j = 1, \dots, N\}. \quad (12)$$

The distribution of  $S$  is conditioned on  $C$ , where the total number  $N = |S|$  and occurrences of the sources are modeled with the Poisson and Bernoulli pdf’s:

$$P(S|C) \propto \frac{\eta^N}{N!} e^{-\eta} \prod_{j=1}^N \rho^{\frac{c(\boldsymbol{\mu}_j)+1}{2}} (1-\rho)^{\frac{1-c(\boldsymbol{\mu}_j)}{2}}, \quad (13)$$

where parameters  $\eta > 0$ ,  $\rho \in (0, 1)$ , and  $c(\boldsymbol{\mu}_j) \in \{-1, 1\}$ .

**Distribution of Agent Intentions** is conditioned on  $S$ , and modeled using the multinomial distribution with parameters  $\boldsymbol{\theta} = [\theta_1, \dots, \theta_j, \dots, \theta_N]$ ,

$$P(R|S) = \prod_{j=1}^N \theta_j^{b_j}, \quad (14)$$

where each  $\theta_j$  is viewed as a prior of selecting  $\mathbf{s}_j \in S$ , and  $\mathbf{s}_j$  is chosen  $b_j$  times to serve as a goal destination,  $b_j = \sum_{i=1}^M \mathbb{1}(r_{ij} = 1)$ ,  $j = 1, \dots, N$ , where  $\mathbb{1}(\cdot)$  denotes the binary indicator function.

**Modeling Agent Behavior:** We make the simplifying assumption that  $R$  and  $Z$  are independent, and that individual intent behavior  $z_i$  is independent from other agents’ behaviors:

$$P(Z) = \prod_{i=1}^M P(z_i), \quad (15)$$

where  $P(z_i)$  is specified for the three types of behavior considered in this paper as follows. The probability that an agent has “single intent” is defined as

$$P(z_i = \text{“single”}) \propto 1 - \kappa, \quad (16)$$

where  $\kappa \in [0, 1]$  is a constant. The probability that an agent has  $n$  “sequential intents” is specified as

$$P(z_i = \text{“sequential”}, n) \propto \kappa^{(n-1)}(1-\kappa), \quad \kappa \in [0, 1]. \quad (17)$$

In our videos, we have  $2 \leq n \leq 3$ . Finally, to define the probability of “change of intent”, we make the assumption that an agent can change the intent only once, and that moment may happen at any time between the start and end of the trajectory with probability  $\gamma \in [0, 1]$ . This is justified in our domain, as explained in Sec. 3. Also, we specify that the new goal can be selected from the remaining  $N - 1$  possible destinations in the scene with a uniform prior distribution. Hence, the probability of “change of intent” is defined as

$$P(z_i = \text{“change”}) = \frac{\gamma}{N-1}. \quad (18)$$

**Trajectory Likelihood:** To address uncertainty about the layout of obstacles, agent’s goal destination(s), and agent’s behavior, we define the likelihood of trajectory  $\Gamma_i$

in terms of the energy that  $a_i$  must spend moving along the trajectory as

$$P(\Gamma_i|C, S, R, z_i) \propto e^{-\lambda \sum_{j, \mathbf{x} \in \Gamma_i} |\vec{F}_{ij}(\mathbf{x}) \cdot \Delta \mathbf{x}|}, \quad (19)$$

where  $\lambda > 0$ , and  $R$  specifies the source(s)  $\mathbf{s}_j$  that  $a_i$  is (sequentially) attracted to. The likelihood in (19) models that when  $a_i$  is far away from  $\mathbf{s}_j$ , the total energy needed to cover that trajectory is bound to be large, and consequently uncertainty about  $a_i$ ’s trajectory is large. Conversely, as  $a_i$  gets closer to  $\mathbf{s}_j$ , uncertainty about the trajectory reduces. Note that applying the principle of Least Action to (19), as in (6), gives the highest likelihood of  $\Gamma_i$ .

## 5 OFFLINE AND ONLINE INFERENCE

Given observations  $\{I, \Gamma'\}$ , we infer the latent variables  $W$  by maximizing the joint posterior defined in (7). We consider offline and online inference.

Offline inference first estimates  $C$ ,  $S$ , and  $R$  over the initial time interval  $(0, t_0)$  observed in the video. These estimates are then used to compute forces  $\{\vec{F}_{ij}(\mathbf{x})\}$  for all agents and their respective goal destinations, as specified in Sec. 3. Finally, the computed forces are used to predict the entire trajectories of agents over  $(t_0, T)$  using the Dijkstra algorithm, and estimate the agents’ intent behaviors.

In online inference, the agent-source relationships  $R^{(t)}$  and trajectories  $\Gamma^{(t)} = \Gamma(0, t)$  are sequentially predicted frame by frame. Thus, new evidence about the agent-source relationships, provided by previous trajectory predictions up to frame  $t$ , is used to re-estimate  $R^{(t+1)}$ . This re-estimation, in turn, is used to predict  $\Gamma^{(t+1)}$  in the next frame  $(t+1)$ .

Note that in offline inference we seek to infer all three types of intent behavior for each agent. In online inference, however, we do not consider “change of intent”, because this would require an explicit modeling of the statistical dependence between  $R$  and  $Z$ , and transition probabilities between  $R^{(t)}$  and  $R^{(t+1)}$  which is beyond our scope.

In the following, we first describe the data-driven MCMC process [4], [3] used for estimating  $C$ ,  $S$ , and  $R$ . Then, we present our sequential estimation of the agents’ trajectories for online inference.

### 5.1 Scene Interpretation

To estimate  $C$ ,  $S$ , and  $R$  over interval  $(0, t_0)$  when the scene in the video can be observed, we use a data-driven MCMC [4], [3], as illustrated in Figures 5 and 6. MCMC provides theoretical guarantees of convergence to the optimal solution.

Each step of our MCMC proposes a new solution  $Y_{\text{new}} = \{C_{\text{new}}, S_{\text{new}}, R_{\text{new}}\}$ . The decision to discard the current solution,  $Y = \{C, S, R\}$ , and accept  $Y_{\text{new}}$  is made based on the acceptance rate,

$$\alpha = \min\left(1, \frac{Q(Y \rightarrow Y_{\text{new}})}{Q(Y_{\text{new}} \rightarrow Y)} \frac{P(Y_{\text{new}}|\Gamma', I)}{P(Y|\Gamma', I)}\right). \quad (20)$$

If  $\alpha$  is larger than a threshold uniformly sampled from  $[0, 1]$ , the jump to  $Y_{\text{new}}$  is accepted. In (20), the proposal

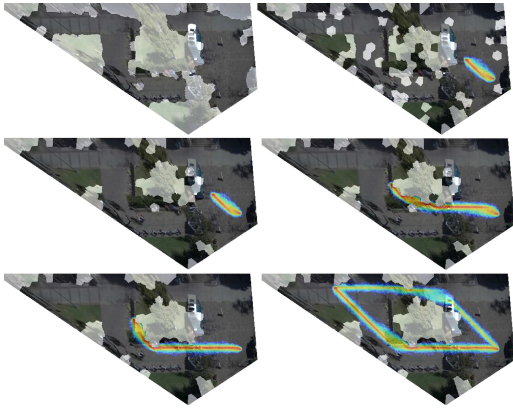


Fig. 5: Top view of the scene from Fig. 2 with the overlaid illustration of the MCMC inference. The rows show the progression of proposals of the constraint map  $C$  in raster scan (the white regions indicate obstacles), and trajectory estimates of agent  $a_i$  with goal to reach  $s_j$  (warmer colors represent higher likelihood  $P(\Gamma_{ij}|C, S, r_{ij} = 1, z_i = \text{“single”})$ ). In the last iteration (bottom right), MCMC estimates that the agent’s goal is to approach  $s_j$  at the top-left of the scene, and finds two equally likely trajectories for this goal.

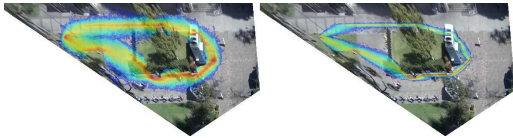


Fig. 6: Top view of the scene from Fig. 2 with the overlaid trajectory predictions of a person who starts at the top-left of the scene, and wants to reach the dark matter in the middle-right of the scene (the food truck). A magnitude of difference in parameters  $\lambda = 0.2$  (on the left) and  $\lambda = 1$  (on the right) used to compute likelihood  $P(\Gamma_{ij}|C, S, R, Z)$  gives similar trajectory predictions. The predictions are getting more certain as the person comes closer to the goal. Warmer colors represent higher likelihood.

distribution is defined as

$$Q(Y \rightarrow Y_{\text{new}}) = Q(C \rightarrow C_{\text{new}})Q(S \rightarrow S_{\text{new}})Q(R \rightarrow R_{\text{new}}), \quad (21)$$

and the posterior distribution

$$P(Y|\Gamma', I) \propto P(C, S, R)P(\Gamma', I|C, S, R, Z'). \quad (22)$$

Each term in (22) is already specified in Sec. 4. Note that in (22) we make the assumption that the observation interval  $(0, t_0)$  is too short for agents to exhibit more complex intent behaviors beyond “single intent”. Therefore, for all agents we set that their initial  $Z' = \{z'_i = \text{“single”}\}$  in  $(0, t_0)$ . Also, note that the prior  $P(Z')$  gets canceled out in the ratio in (20), and hence  $P(Z')$  is omitted in (22).

The initial  $C$  is proposed by setting  $c(\mathbf{x}) = 1$  at all locations covered by observed trajectories  $\Gamma'$ , and randomly setting  $c(\mathbf{x}) = -1$  or  $c(\mathbf{x}) = 1$  for other locations. The initial number  $N$  of sources in  $S$  is probabilistically sampled from the Poisson distribution of (13), while their

layout is estimated as  $N$  most frequent stopping locations in  $\Gamma'$ . Given  $\Gamma'$  and  $S$ , we probabilistically sample the initial  $R$  using the multinomial distribution in (14). In the subsequent MCMC iterations, new solutions  $C_{\text{new}}$ ,  $S_{\text{new}}$ , and  $R_{\text{new}}$  are sequentially proposed and accepted as current solutions based on the acceptance rate  $\alpha$ .

**The Proposal of  $C_{\text{new}}$**  randomly chooses  $\mathbf{x} \in \Lambda$ , and reverses its polarity,  $c_{\text{new}}(\mathbf{x}) = -c(\mathbf{x})$ . The proposal distribution  $Q(C \rightarrow C_{\text{new}}) = Q(c_{\text{new}}(\mathbf{x}))$  is data-driven.  $Q(c_{\text{new}}(\mathbf{x}) = 1)$  is defined as the normalized average speed of people observed at  $\mathbf{x}$ , and  $Q(c_{\text{new}}(\mathbf{x}) = -1) = 1 - Q(c_{\text{new}}(\mathbf{x}) = 1)$ .

**The Proposal of  $S_{\text{new}}$**  includes the “death” and “birth” jumps. The birth jump randomly chooses  $\mathbf{x} \in \Lambda_1$ , and adds a new source  $s_{N+1} = (\mu_{N+1}, \Sigma_{N+1})$  to  $S$ , resulting in  $S_{\text{new}} = S \cup \{s_{N+1}\}$ , where  $\mu_{N+1} = \mathbf{x}$ , and  $\Sigma_{N+1} = \text{size}^2 \mathcal{I}$ , where size is the scene size (in pixels). The death jump randomly chooses an existing source  $s_j \in S$ , and removes it from  $S$ , resulting in  $S_{\text{new}} = S \setminus \{s_j\}$ . The ratio of the proposal distributions is specified as  $\frac{Q(S \rightarrow S_{\text{new}})}{Q(S_{\text{new}} \rightarrow S)} = 1$ , indicating no preference to either “death” or “birth” jumps. That is, the proposal of  $S_{\text{new}}$  is governed by the Poisson prior of (13), and trajectory likelihoods  $P(\Gamma'|C, S, R, Z')$ , given by (19), when computing the acceptance rate  $\alpha$ .

**The Proposal of  $R_{\text{new}}$**  randomly chooses one person  $a_i \in A$  with goal  $s_j$ , and performs one of the three possible actions: (i) randomly changes  $a_i$ ’s goal to  $s_k \in S$ , (ii) randomly adds a new goal  $s_k \in S$  to  $a_i$  if  $\sum_j r_{ij} < n$  where  $n$  is the maximum number of goals (in our domain  $n = 3$ ), and (iii) randomly removes one of current goals of  $a_i$  if  $\sum_j r_{ij} > 1$ . The changes in the corresponding relationships  $r_{ij} \in R$  result in  $R_{\text{new}}$ . The ratio of the proposal distributions is  $\frac{Q(R \rightarrow R_{\text{new}})}{Q(R_{\text{new}} \rightarrow R)} = 1$ . This means that the proposal of  $R_{\text{new}}$  is governed by the multinomial prior  $P(R|S)$  and likelihoods  $P(\Gamma'|C, S, R, Z')$ , given by (14) and (19), when computing  $\alpha$  in (20).

Importantly, the random proposals of  $R_{\text{new}}$  ensure that for every agent  $a_i$  we have  $\sum_j r_{ij, \text{new}} \geq 1$ . Since our assumption is that in  $(0, t_0)$  the agents may have only a single intent, we consider that agent  $a_i$  first wants to reach the closest source  $s_j$  for which  $r_{ij, \text{new}} = 1$ ,  $\Gamma'_i = \Gamma'_{ij}$ , out of all other sources  $s_k$  that also have  $r_{ik, \text{new}} = 1$  and which the agent can visit later after time  $t_0$ . This closest source  $s_j$  is then used to compute the likelihood  $P(\Gamma'_{ij}|C, S, r_{ij} = 1, z'_i = \text{“single”})$ , as required in (20), and thus conduct the MCMC jumps.

## 5.2 Offline Inference of $\Gamma$ and $Z$

From the MCMC estimates of  $C, S, R$ , we readily estimate forces  $\{\vec{F}_{ij}\}$ , given by (5). Then we proceed to predicting trajectories  $\{\Gamma_i\}$  and intent behavior  $\{z_i\}$  in the future interval  $(t_0, T)$ .

**The single-intent case.** When the MCMC estimates that  $a_i$  has only a single goal destination,  $\sum_j r_{ij} = 1$ , we readily predict  $z_i = \text{“single”}$ , and estimate a globally optimal trajectory  $\Gamma_i = \Gamma_{ij}$  using the Dijkstra algorithm. The Dijkstra path ends at location  $\mathbf{x}_j$  of source  $s_j$  for which



$r_{ij} = 1$ , where  $a_i$  is taken to remain still until the end of the time horizon  $T$ .

**The sequential and change of intent cases.** When the MCMC estimates that  $\sum_j r_{ij} > 1$ , we hypothesize that the agent could have either the “sequential” or “change” of intent behavior. In this case, we jointly estimate the optimal  $(\Gamma_i, z_i)^*$  pair by maximizing their joint likelihood:

$$(\Gamma_i, z_i)^* = \arg \max_{\Gamma_i, z_i} P(\Gamma_i | C, S, R, z_i) P(z_i), \quad (23)$$

for all possible configurations of  $(\Gamma_i, z_i)$ , where  $z_i \in \{\text{“sequential”}, \text{“change”}\}$  and  $\Gamma_i$  is the Dijkstra path spanning a particular sequence of selected sources from  $S$ , as explained in more detail below.

Let  $J = \{j : r_{ij} = 1, j = 1, \dots, N\}$  denote indices of the sources from  $S$  that MCMC selected as goal destinations for  $a_i$ . Also, recall that for our videos of public spaces, it is reasonable to expect that people may have a maximum of  $|J| \leq n = 3$  intents in interval  $(0, T)$ . The relatively small size of  $J$  allows us to exhaustively consider all permutations of  $J$ , where each permutation uniquely identifies the Dijkstra path  $\Gamma_i$  between  $a_i$ 's location at time  $t_0$ ,  $\mathbf{x}(t_0)$  and the sequence of locations  $\mathbf{x}_j$ ,  $j \in J$ , that  $a_i$  visits along the trajectory. For each permutation of  $J$ , and each intent behavior  $z_i \in \{\text{“sequential”}, \text{“change”}\}$ , we compute the joint likelihood given by (23), and infer the maximum likelihood pair  $(\Gamma_i, z_i)^*$ .

### 5.3 Online Inference of $\Gamma^{(t)}$ and $R^{(t)}$

In online inference, we first estimate  $C$ ,  $S$ , and  $R^{(t_0)}$  over the time interval  $(0, t_0)$  using the data-driven MCMC as described in Sec. 5.1. Then, we compute forces  $\{\bar{F}_{ij}^{(t_0)}(\mathbf{x})\}$  for all agents and their respective goal destinations identified in  $R^{(t_0)}$ . Also, we take the observed trajectories  $\Gamma'$  as initial trajectory estimates,  $\Gamma_i^{(t_0)} = \Gamma_i(0, t_0) = \Gamma'_i$ ,  $i = 1, \dots, M$ . This initializes our online inference of  $\Gamma^{(t)}$  and  $R^{(t)}$  at times  $t \in (t_0, T)$ . Then,  $\Gamma^{(t)}$  are taken to provide new cues for predicting  $R^{(t+1)}$  using the MCMC described in Sec. 5.1. Any updates in  $R^{(t+1)}$  are used to compute  $\{\bar{F}_{ij}^{(t+1)}(\mathbf{x})\}$ , and subsequently update  $\Gamma^{(t+1)}$  as follows.

Recall that in online inference we do not consider “change of intent” behavior. In case  $R^{(t)}$  estimates that agent  $a_i$  has more than one goal,  $\sum_j r_{ij}^{(t)} \geq 1$ , our online predictions of  $\Gamma_i^{(t)}$  use the heuristic that  $a_i$  visits the goal destinations identified in  $R^{(t)}$  in the order of how far they are from the current location of  $a_i$ . Consequently, predictions of  $\Gamma_i^{(t)}$  become equivalent for both “single” and “sequential” intent behavior, since  $a_i$  always wants to reach the closest goal destination first, i.e.,  $\Gamma_i^{(t)} = \Gamma_{ij}^{(t)}$  for  $r_{ij}^{(t)} = 1$  and  $\mathbf{x}_j$  is the closest location to  $a_i$  at time  $t$ . From (19), it is straightforward to derive the conditional likelihood of  $\Gamma_i^{(t+1)}$ , given  $\Gamma_i^{(t)}$  and  $R^{(t)}$  as

$$P(\Gamma_{ij}^{(t+1)} | C, S, r_{ij}^{(t)} = 1, z_i = \text{“single”}, \Gamma_{ij}^{(t)}) \propto e^{-\lambda(|\bar{F}_{ij}^{(t)}| \cdot |\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}| + \min_{\mathbf{x}} \sum_{\mathbf{x}=\mathbf{x}^{(t+1)}}^{\mathbf{x}^j} |\bar{F}_{ij}^{(t)}(\mathbf{x}) \cdot \Delta \vec{\mathbf{x}}|)}, \quad (24)$$

where the second term in (24) represent the energy that  $a_i$  needs to spend while walking along the Dijkstra path from  $\mathbf{x}^{(t+1)}$  to the goal destination  $\mathbf{x}_j$ .

Our online inference is summarized below.

- **Input:** Observed trajectories  $\Gamma' = \Gamma'(0, t_0) = \Gamma^{(t_0)}$ , the MCMC estimates of  $C$ ,  $S$ ,  $R^{(t_0)}$  computed as described in Sec. 5.1, and time horizon  $T$ .
- **Online trajectory prediction:** For every  $a_i$  identify the closest goal destination  $r_{ij}^{(t)} = 1$ , and compute  $\Gamma_{ij}^{(t+1)} = [\Gamma_{ij}^{(t)}, \mathbf{x}^{(t+1)}]$ , where the next location  $\mathbf{x}^{(t+1)}$  of the trajectory is estimated as an average of probabilistic samples  $\xi$  generated from the conditional likelihood of (24):
 
$$\mathbf{x}^{(t+1)} = \text{MEAN}(\xi),$$

$$\xi \sim e^{-\lambda(|\bar{F}_{ij}^{(t)}| \cdot |\xi - \mathbf{x}^{(t)}| + \min_{\mathbf{x}} \sum_{\mathbf{x}=\xi}^{\mathbf{x}^j} |\bar{F}_{ij}^{(t)}(\mathbf{x}) \cdot \Delta \vec{\mathbf{x}}|)}, \quad (25)$$
- **Stopping criterion:** If  $\Gamma_{ij}^{(t+1)}$  exists out of the scene, agent  $a_i$  has visited all goal destinations identified in  $R^{(t)}$ , or time reaches the horizon  $t = T$ .

## 6 RESULTS

For evaluation, we use toy examples and real outdoor scenes. We present six types of results: (a) localization of “dark matter”  $S$ , (b) estimation of human intents  $R$ , (c) prediction of human trajectories  $\Gamma$ , (d) inference of “single”, “sequential”, and “change” intent behavior, and (e) functional object clustering. These results are computed on unobserved video parts, given access to an initial part of the video footage. We study how our performance varies as a function of the length of the observed footage. For real scenes, note that annotating ground truth of non-walkable surfaces  $C$  in a scene is difficult, since human annotators provide inconsistent subjective estimates (e.g., grass lawn can be truly non-walkable in one part of the scene, but walkable in another). Therefore, we do not quantitatively evaluate our inference of  $C$ .

Note that our evaluations (a)–(d) significantly extend the work of [26] which presents results on only detecting “exits” and “vehicles” as “dark matter” in the scene, and predicting human trajectories for “single intent” that are bound to end at locations of “exits” and “vehicles”. A comparison of our results for (a) with existing approaches to object detection would be unfair, since we do not have access to annotated training examples of the objects as most appearance-based methods for object recognition.

**Evaluation Metrics:** For evaluating our trajectory prediction, we compute a modified Hausdorff distance (MHD) between the ground-truth trajectory  $\Gamma$  and predicted trajectory  $\Gamma^*$  as

$$\begin{aligned} \text{MHD}(\Gamma, \Gamma^*) &= \max(d(\Gamma, \Gamma^*), d(\Gamma^*, \Gamma)), \\ d(\Gamma, \Gamma^*) &= \frac{1}{|\Gamma|} \sum_{\mathbf{x} \in \Gamma} \min_{\mathbf{x}^* \in \Gamma^*} |\mathbf{x} - \mathbf{x}^*|. \end{aligned} \quad (26)$$

For comparison with [26], we also compute the negative log-likelihood (NLL),  $\log P(\Gamma_{ij} | \cdot)$ , of the ground-truth trajectory  $\Gamma_{ij}(t_1, t_2) = \{\mathbf{x}(t_1) = \mathbf{x}_i, \dots, \mathbf{x}(t_2) = \mathbf{x}_j\}$ . From

(19), NLL can be expressed as

$$\text{NLL}_P(\Gamma) = \frac{\lambda}{t_2 - t_1} \sum_{t=t_1}^{t_2-1} |\vec{F}_{ij}(\mathbf{x}(t)) \cdot (\mathbf{x}(t+1) - \mathbf{x}(t))|, \quad (27)$$

where  $\vec{F}_{ij}(\mathbf{x}(t))$  is our estimate of the force field affecting the  $i$ th person in the video.

For evaluating localization of “dark matter”,  $S$ , we use the standard overlap criterion, i.e., the intersection-over-union ratio  $IOU$  between our detection and ground-truth bounding box around the functional object. True positive detections are estimated for  $IOU \geq 0.5$ .

For evaluating human-destination relationships,  $R$ , we compute a normalized Hamming distance between the ground-truth  $r_{ij}$  and predicted  $r_{ij}^*$  binary indicators of human-destination relationships,  $\frac{\sum_j \mathbb{1}(r_{ij}=r_{ij}^*)}{\sum_j r_{ij}}$ .

Finally, for evaluating intent behavior, we use the standard classification recall and precision estimated from the confusion matrix of the three intent-behavior classes.

**Baselines:** For conducting an ablation study, we evaluate the following baselines and compare them with our full approach in the offline inference setting. The first three baselines are evaluated only on trajectories where people truly have a “single” intent, whereas the sixth baseline is evaluated on all trajectories. (1) “Shortest path” (SP) estimates the trajectory as a straight line, disregarding obstacles in the scene, given the MCMC estimates of  $S$  and  $R$ . SP does not infer latent  $C$ , and in this way tests the influence of estimating  $C$  on our overall performance. (2) “Random Walk” (RW) sequentially predicts the trajectory frame by frame, given the MCMC estimates of  $C$  and  $S$ , where every prediction randomly selects one destination  $j$  from  $S$  and prohibits landing on non-walkable areas. RW does not estimate  $R$ , and in this way tests the influence of estimating  $R$  on our overall performance. (3) “Lagrangian Physical Move” (PM) predicts the trajectory, given estimates of  $C$  and  $S$ , under the sum of forces from all sources,  $\vec{F}_{\text{classic}}(\mathbf{x}) = \sum_j \vec{F}_{ij}(\mathbf{x}) + \vec{F}^-(\mathbf{x})$ , as defined in Section 3 for the Lagrangian Mechanics. As RW, PM does not estimate  $R$ . (4) “Greedy move” (GM) makes the assumption that every person wants to go to the initially closest functional object, and thus sequentially predicts both the trajectory  $\Gamma_i^{(t)}$  and destination  $r_{ij}^{(t)}$  frame by frame, given the MCMC estimates of  $C$  and  $S$ , where the latter is estimated by maximizing the following likelihood:

$$j^* = \arg \max_j P(r_{ij} | \Gamma_i^{(t)}) \propto e^{\tau(|\mathbf{x}_j - \mathbf{x}^{(t)}| - |\mathbf{x}_j - \mathbf{x}_i|)}. \quad (28)$$

where  $\mathbf{x}^{(t)}$  is the last location of  $\Gamma_i^{(t)}$ . This baseline also tests the merit of our MCMC estimation of  $R$ .

**Comparison with Related Approaches.** We are not aware of prior work on estimating  $S$ ,  $R$ , and  $Z$  in the scene without access to manually annotated training labels of objects. We compare only with the state of the art method for trajectory prediction [26].

**Input Parameters.** In our default setting, we consider the first 50% of trajectories as visible, and the remainder as unobserved. We use the following model parameters:  $\beta =$

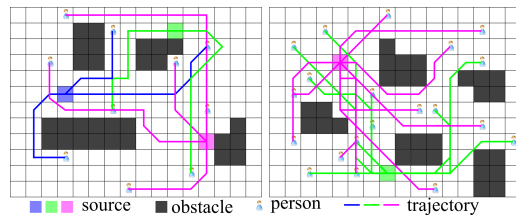


Fig. 7: The ground truth of two scenes in the toy dataset. There are three “dark-matter” objects in each scene represented by colored squares whereas the black regions are defined as obstacles. The trajectories are created by i) assigning an agent with a starting position and one of the objects as destination, and ii) sampling a path between the starting position and the function object. The colors of the trajectories indicate the corresponding “dark matter”.

S	S&R				NLL			
	10	20	50	100	10	20	50	100
2	0.95	0.97	0.96	0.96	1.35	1.28	1.17	1.18
3	0.87	0.90	0.94	0.94	1.51	1.47	1.35	1.29
5	0.63	0.78	0.89	0.86	1.74	1.59	1.36	1.37
8	0.43	0.55	0.73	0.76	1.97	1.92	1.67	1.54

TABLE 3: Accuracy of  $S$  and  $R$  averaged over all agents, and NLL on the toy dataset.  $S\&R$  is a joint accuracy, where the joint  $S\&R$  is deemed correct if both  $S$  and  $R$  are correctly inferred for every agent. The first column lists the number of sources  $|S|$ , and the second row lists the number of agents  $|A|$ .

.05,  $\lambda = 0.5$ ,  $\rho = 0.95$ . From our experiments, varying these parameters in intervals  $\beta \in [0.01, .1]$ ,  $\lambda \in [0.1, 1]$ , and  $\rho \in [0.85, 0.98]$  does not change our results, suggesting that we are relatively insensitive to the specific choices of  $\beta, \lambda, \rho$  over large intervals.  $\eta$  is known.  $\theta$  and  $\psi$  are fitted from observed data.

## 6.1 Toy Dataset

The toy dataset allows us to methodologically test our approach with respect to each dimension of the scene complexity, while fixing the other parameters. The scene complexity is defined in terms of the number of agents in the scene and the number of sources. All agents are taken to have only “single” intent. The scene parameters are varied to synthesize the toy artificial scenes in a rectangle random layout, where the ratio of obstacle pixels over all pixels is about 15%. We vary  $|S|$  and  $|A|$ , and we generate 3 random scene layouts for each setting of parameters  $|S|$  and  $|A|$ . Fig. 7 shows two examples from our toy dataset. For inference, we take the initial 50% of trajectories as observed. Tab. 3 shows that our approach can handle large variations in each dimension of the scene complexity.

## 6.2 Real Scenes

### 6.2.1 Datasets

We use 8 different real scenes for the experiments: ① Courtyard dataset [6]; video sequences of two squares ② SQ1 and ③ SQ2 annotated by VATIC [49]; ④ VIRAT ground dataset [5]; new scenes including CourtyardNew

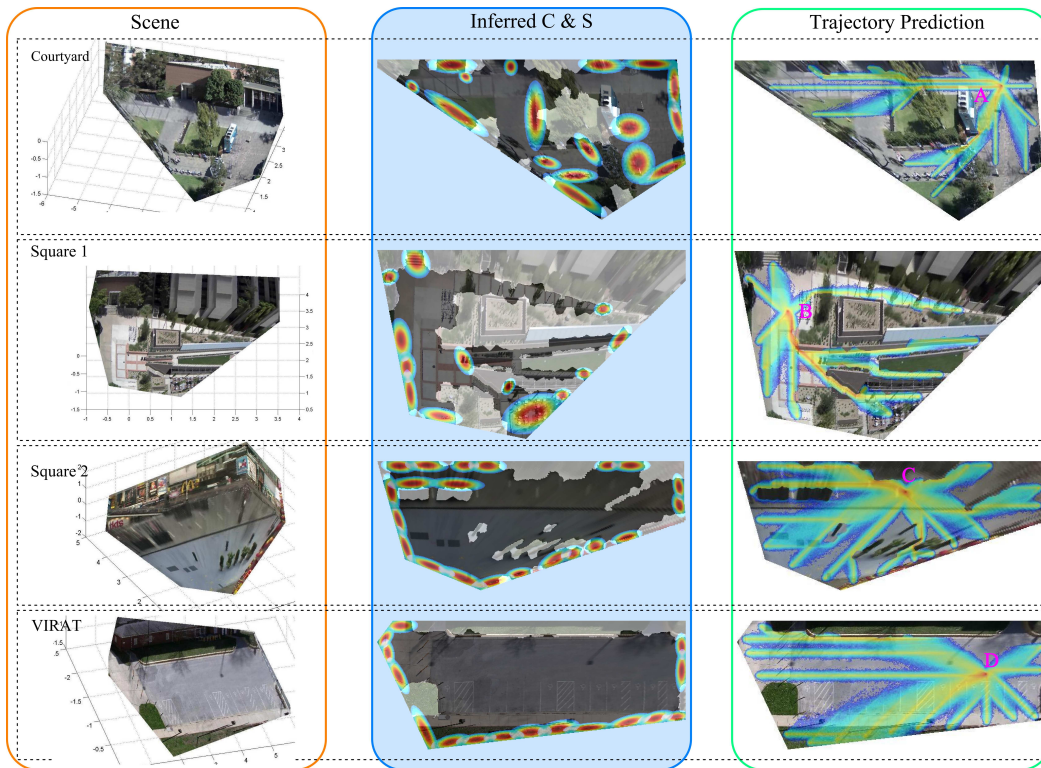


Fig. 8: Qualitative experiment results for “dark matter” localization and “single intent” prediction in 4 scenes. Each row is one scene. The 1st column is the reconstructed 3D surfaces of each scene. The 2nd column is the estimated layout of obstacles (the white masks) and dark matter (the Gaussians). The 3rd column shows the trajectory prediction by sampling. We predict the future trajectory for an agent at some position ( $A, B, C, D$ ) in the scene toward each potential source in  $S$ . The warm and cold colors represent high and low probability of visiting the positions respectively. Note that we cropped and projected the ground surface onto the top-down view which results into the irregular polygons.

Dataset	$ S $	Source Name
① Courtyard	19	bench/chair, food truck, bldg, vending machine, trash can, exit
② SQ1	15	bench/chair, trash can, bldg, exit
③ SQ2	22	bench/chair, trash can, bldg, exit
④ VIRAT	17	vehicle, exit
⑤ CourtyardNew	16	bench/chair, exit, bldg
⑥ AckermanUnion1	16	table, trash can, bldg, exit
⑦ AckermanUnion2	16	bench/chair, bldg, exit
⑧ AerialVideo	16	table, vehicle, trash can, bldg, exit

TABLE 4: Summary of “dark matter” for the datasets

⑤, AckermanUnion1 ⑥ and AckermanUnion2 ⑦; AerialVideo ⑧ from UCLA Aerial Event dataset [7]. SQ1 is 20min,  $800 \times 450$ , 15 fps. SQ2 is 20min,  $2016 \times 1532$ , 12 fps. We use the same scene A of VIRAT as in [26]. New videos ⑤⑥⑦ last 2min, 30min and 30min respectively. We select video 59 from [7]. The last four videos all have 15 fps and  $1920 \times 1080$  resolution. For “single intent” prediction, we allow the initial observation of 50% of the video footage, which for example gives about 300 trajectories in ①.

While the ground-truth annotation of “single” and “sequential” intent behaviors is straightforward, in real scenes, we have encountered a few ambiguous cases of “change of intent” where different annotators have disagreed about “single” or “change of intent” behavior for the same trajectory. In such ambiguous cases we used a majority vote

Dataset	S		R		NLL		MHD			
	Our	Initial	Our	GM	Our	[26] RW	Our	RW	SP	PM
①	<b>0.89</b>	0.23	<b>0.52</b>	0.31	<b>1.635</b>	- 2.197	<b>14.8</b>	243.1	43.2	207.5
②	<b>0.87</b>	0.37	<b>0.65</b>	0.53	<b>1.459</b>	- 2.197	<b>11.6</b>	262.1	39.4	237.9
③	<b>0.93</b>	0.26	<b>0.49</b>	0.42	<b>1.621</b>	- 2.197	<b>21.5</b>	193.8	27.9	154.2
④	<b>0.95</b>	0.25	<b>0.57</b>	0.46	<b>1.476</b>	1.594 2.197	<b>16.7</b>	165.4	21.6	122.3
⑤	<b>0.81</b>	0.19	<b>0.75</b>	0.50	<b>0.243</b>	- 2.197	<b>20.1</b>	71.32	27.5	181.7
⑥	<b>0.81</b>	0.38	<b>0.63</b>	0.42	<b>0.776</b>	- 2.197	<b>13.7</b>	183.0	17.5	128.6
⑦	<b>0.88</b>	0.32	<b>0.56</b>	0.39	<b>1.456</b>	- 2.197	<b>19.3</b>	150.1	26.9	119.2
⑧	<b>0.63</b>	0.25	<b>0.67</b>	0.33	<b>1.710</b>	- 2.197	<b>15.9</b>	289.3	26.5	137.0

TABLE 5: Qualitative results of “single intent”

Obs. %	S	R	NLL	MHD
45%	0.85	0.47	1.682	15.7
40%	0.79	0.41	1.753	16.2

TABLE 6: Results on ① with different observed ratios

as ground truth for intent behavior.

### 6.2.2 Experiment 1: Functional Object Localization and Single Intent Prediction

In this experiment, we use only “single intent” trajectories of real-world datasets. Our qualitative results are shown in Figure 8, and quantitative evaluation is presented in Tab. 5. As can be seen: (1) We are relatively insensitive to the specific choice of model parameters. (2) We handle challenging scenes with arbitrary layouts of dark matter, both in the middle of the scene and at its boundaries. From Tab. 5, the comparison with the baselines demonstrates that the initial

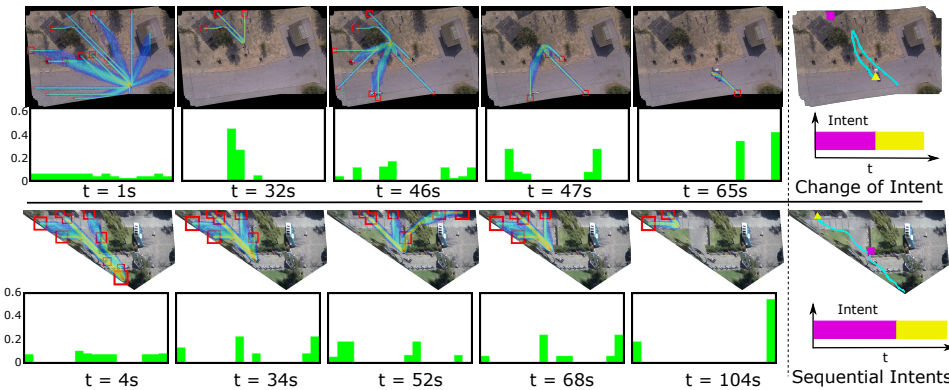
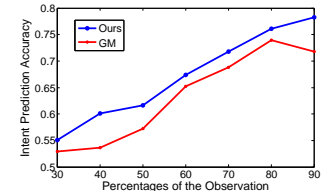
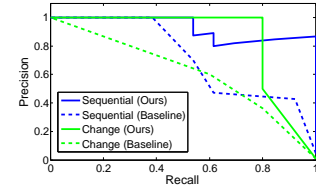


Fig. 9: Qualitative result of “sequential intents” and “change of intent”. Left: online prediction where the red bounding boxes represent the possible intents at a certain moment and a larger bounding box indicates an intent with higher probability; histograms represent the probabilities of each functional object being the intent of the agent at a given moment. Right: offline intent types and intents inference based on the full observation of trajectories, where the square is the first intent and the triangle is the second intent.



(a) Online intent prediction accuracy.



(b) Offline intent type inference.

Fig. 10: PR curves of “sequential intents” and “change of intent”.

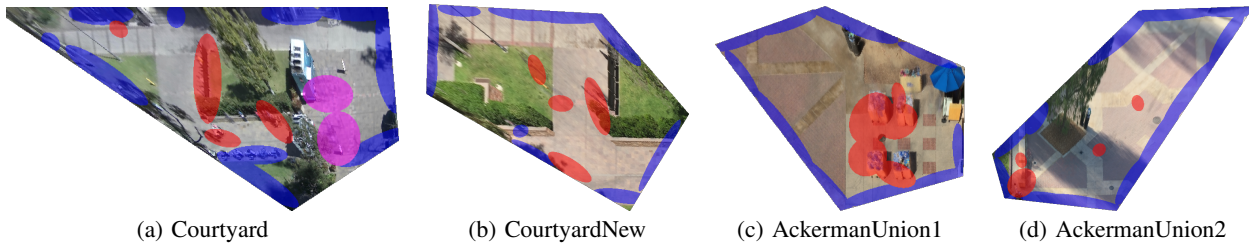


Fig. 11: Functional object clustering results. Each ellipse map is an inferred functional object and the four different colors (i.e., magenta, red, blue, and yellow) represent four latent functional classes.

guess of sources based on partial observations gives very noisy results. These noisy results are significantly improved in our DD-MCMC inference. Also, our method is a slightly better than the baseline GM if there are a few obstacles in the middle of the scene. But we get a huge performance improvement over GM if there are complicated obstacles in the scene. This shows that our global plan based relation prediction is better than GM.

Based on  $S$  and  $C$ , we model human motion probabilistically given their goals and understanding of scenes so that we can predict their future trajectories with probability. The prediction accuracy on the four scenes are summarized in Table 5. It appears that our results are also superior to the random walk. The baselines RW and PM produce bad trajectory prediction. While SP yields good results for scenes with a few obstacles, it is brittle for more complex scenes which we successfully handle. When the size of  $S$  is large (e.g., many exists from the scene), our estimation of human goals may not be exactly correct. However, in all these error cases, the goal that we estimate is not spatially far away from the true goal. Also, in these cases, the predicted trajectories are also not far away from the true trajectories measured by MHD and NLL. Our performance downgrades gracefully with the reduced observation time as Table 6 indicates. We outperform the state of the art [26]. Note that the MHD absolute values produced by

our approach and [26] are not comparable, because this metric is pixel based and depends on the resolution of reconstructed 3D surface.

Our results show that our method successfully addresses surveillance scenes of various complexities.

### 6.2.3 Experiment 2: “Sequential Intents” and “Change of Intent” Inference

For the evaluation, we randomly select 500 trajectories from all scenes and manually annotate the types of intent for each of them. In all these trajectories, there are 13 “sequential intents” instances, 5 “change of intent” instances, and the remaining trajectories all have “single intent”.

The qualitative results of both online prediction and offline inference are visualized in Fig. 9. Note that we automatically infer the vehicles as “dark matter” in (8). For the “change of intent” case, the agent switched the intent from a building to a vehicle at around 46s; for the “sequential intents” case, the agent first went a trash can near the bench (latent behind the bush) to throw trash (52s), and then left for one of the exits (68s). It appears that i) our online predictions can reflect the intent changes of the observed agents and ii) with the full trajectory observation, the offline inference correctly recognizes the intent types and the corresponding temporal passing of the intents.

The online intent prediction accuracy is shown in Fig. 10a, which indicates that ours consistently outperform

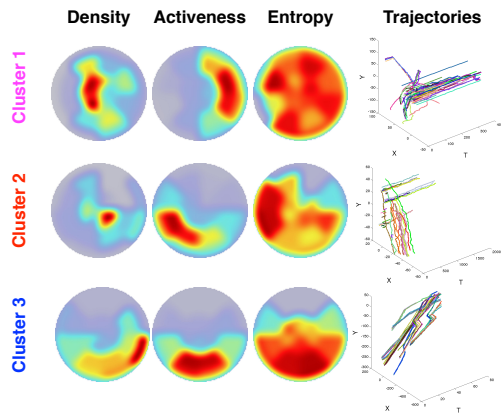


Fig. 12: Mean feature maps of the three clusters of “dark matter” and the associated trajectories of an example functional object within each cluster. All objects of the same cluster are aligned to a reference direction.

the baseline. In Fig. 10b, we plot the precision-recall curves for the offline intent type inference. The baseline method uses a motion-based measurement of the trajectory, i.e., the longest stationary time before arriving the intent for the “sequential intents”, and the largest turning angle for the “change of intent”. Ours yields very high mAPs for both “sequential intents” and “change of intent” (0.93 and 0.80 respectively) than the baseline (0.66 and 0.43 respectively).

#### 6.2.4 Experiment 3: Functional Objects Clustering

The estimated trajectories of people can be viewed as a summary of their behavior under the attraction and repulsion forces of “dark matter” in the scene. We show that it is possible to cluster objects (of different semantic classes) by the properties of human behavior, and in this way *discover* different types of “dark matter”.

For such clustering we use the following features characterizing the nearby region of each identified “dark matter”: a) the density of the associated agents around “dark matter”, i.e., a density map; b) the spatial distribution of the average velocity magnitude of the associated agents, i.e., an activeness map; c) the spatial distribution of the entropy of moving directions of the associated agents, i.e., an entropy map. We assume that all agents have a single intent and associate their trajectories with their intents for computing the features. Similar to the shape context features [50], we convert the three maps into three histograms and concatenate them into a feature vector. Based on the feature vectors, we then perform K-means clustering to group the inferred “dark matter”. Note that feature maps are aligned within the same cluster by rotation and mirroring.

We cluster all objects in the real scenes into 3 types of “dark matter”. Fig. 11 shows the clustering result in 4 scenes and Fig. 12 visualizes the mean feature maps of each cluster and the associated trajectories around a example functional object in each cluster. The visualized results confirm that without the appearance and geometry information of the “dark matter”, we are able to discover meaningful functional classes by analyzing human behaviors, which show clear semantic meanings: i) magenta regions: queuing

areas; ii) red regions: areas where people stand or sit for a long time (e.g., benches, chairs, tables, vending machines); iii) blue regions: exits or buildings. Note that sometimes there are a few small objects in the scenes, e.g., trash cans in Fig. 11c, that are not identified as “dark matter” by our approach, simply because they were never used by the agents in the videos. Interestingly, we occasionally obtain one to two red regions around the lawn (e.g. Fig. 11b) or on the square (e.g. Fig. 11d) since multiple agents were standing there for a long time.

## 7 CONCLUSION

We have addressed the problem of predicting human trajectories in unobserved parts of videos of public spaces, given access only to an initial excerpt of the videos in which most of the human trajectories have not yet reached their respective destinations. We have formulated this problem as that of reasoning about latent human intents to approach “dark matter” in the scene, and, consequently, identifying a functional map of the scene. Our work extends the classical Lagrangian mechanics to model the scene as a physical system wherein: i) “dark matter” exerts attraction forces on people’s motions, and ii) people are viewed as agents who can have intents to approach particular “dark matter”. For evaluation we have used the benchmark VIRAT, UCLA Courtyard and UCLA Aerial Event datasets, as well as our five videos of public spaces. We have shown that it is possible to cluster objects of different semantic classes by the properties of human motion behavior in their surrounding, and in this way discover different types of “dark matter”. One limitation of our method is that it does not account for social interactions between agents, which seems a promising direction for future work.

## ACKNOWLEDGMENT

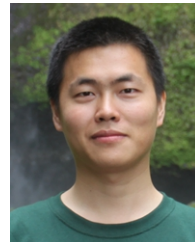
This work is supported by grants NSF IIS 1423305, ONR MURI N00014-16-1-2007, and a DARPA SIMPLEX project N66001-15-C-4035.

## REFERENCES

- [1] H. Pirsivash, D. Ramanan, and C. C. Fowlkes, “Globally-optimal greedy algorithms for tracking a variable number of objects,” in *CVPR*, 2011.
- [2] Y. Zhao and S.-C. Zhu, “Image parsing via stochastic scene grammar,” in *NIPS*, 2011.
- [3] Z. Tu and S.-C. Zhu, “Image segmentation by data-driven markov chain monte carlo,” *TPAMI*, vol. 24, no. 5, pp. 657–673, 2002.
- [4] J. Kwon and K. M. Lee, “Wang-Landau monte carlo-based tracking methods for abrupt motions,” *TPAMI*, vol. 35, no. 4, pp. 1011–1024, 2013.
- [5] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-c. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsivash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai, “A large-scale benchmark dataset for event recognition in surveillance video,” in *CVPR*, 2011.
- [6] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu, “Cost-sensitive top-down / bottom-up inference for multiscale activity recognition,” in *ECCV*, 2012.
- [7] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, “Joint inference of groups, events and human roles in aerial videos,” in *CVPR*, 2015.

- [8] J. Gall, A. Fossati, and L. V. Gool, "Functional categorization of objects using real-time markerless motion capture," in *CVPR*, 2011.
- [9] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: using spatial and functional compatibility for recognition," *TPAMI*, vol. 31, no. 10, pp. 1775–1789, Oct. 2009.
- [10] V. Delaitre, D. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. Efros, "Scene semantics from long-term observation of people," in *ECCV*, 2012.
- [11] A. Pieropan, C. Ek, and H. Kjellstrom, "Functional object descriptors for human activity modeling," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2013.
- [12] P. Wei, Y. Zhao, N. Zheng, and S.-C. Zhu, "Modeling 4d human-object interactions for event and object recognition," in *ICCV*, 2013.
- [13] H. Grabner, J. Gall, and L. V. Gool, "What makes a chair a chair?" in *CVPR*, 2011.
- [14] M. W. Turek, A. Hoogs, and R. Collins, "Unsupervised learning of functional categories in video scenes," in *ECCV*, 2010.
- [15] S. Oh, A. Hoogs, M. Turek, and R. Collins, "Content-based retrieval of functional objects in video using scene context," in *ECCV*, 2010.
- [16] B. Qin, Z. Chong, T. Bandyopadhyay, M. Ang, E. Frazzoli, and D. Rus, "Learning pedestrian activities for semantic mapping," in *ICRA*, 2014.
- [17] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE TCSVT*, 2008.
- [18] A. Abrams, J. Tucek, J. Little, N. Jacobs, and R. Pless, "LOST: Longterm observation of scenes (with tracks)," in *Workshop on Applications of Computer Vision (WACV)*, 2012, (oral presentation).
- [19] S. Ali and M. Shah, "A Lagrangian particle dynamics approach for crowd flow segmentation and stability analysis," in *CVPR*, 2007.
- [20] B. Solmaz, B. E. Moore, and M. Shah, "Identifying behaviors in crowd scenes using stability analysis for Dynamical Systems," *TPAMI*, vol. 34, no. 10, pp. 2064–2070, Oct. 2012.
- [21] H. Gong, J. Sim, M. Likhachev, and J. Shi, "Multi-hypothesis motion planning for visual object tracking," in *ICCV*, 2011.
- [22] J. Barraquand, B. Langlois, and J.-C. Latombe, "Numerical potential field techniques for robot path planning," *TSMC*, vol. 22, no. 2, 1992.
- [23] W. Shao and D. Terzopoulos, "Autonomous pedestrians," in *SCA*, 2005.
- [24] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–49, Dec. 2009.
- [25] B. D. Ziebart, N. Ratliff, G. Gallagher, C. Mertz, K. Peterson, J. A. Bagnell, M. Hebert, A. K. Dey, and S. Srinivasa, "Planning-based prediction for pedestrians," in *IROS*, 2009.
- [26] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *ECCV*, 2012.
- [27] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, pp. 4282–4286, 1995.
- [28] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *CVPR*, 2009.
- [29] E. Pellegrini and v. G. Schindler, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *ICCV*, 2009.
- [30] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social Istm: Human trajectory prediction in crowded spaces," in *CVPR*, 2016.
- [31] H. S. Park, J.-J. Hwang, Y. Niu, and J. Shi, "Egocentric future localization," in *CVPR*, 2016.
- [32] B. Zhou, X. Wang, and X. Tang, "Understanding collective crowd behaviors: Learning a Mixture model of Dynamic pedestrian-Agents," in *CVPR*, 2012.
- [33] C. Baker, J. Tenenbaum, and R. Saxe, "Goal inference as inverse planning," in *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*, 2007.
- [34] Y. Kong, D. Kit, and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *ECCV*, 2014.
- [35] T. Lan, T.-C. Chen, and S. Savarese, "A hierarchical representation for future action prediction," in *ECCV*, 2014.
- [36] D.-A. Huang and K. M. Kitani, "Action-reaction: Forecasting the dynamics of human interaction," in *ECCV*, 2014.
- [37] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV*, 2011.
- [38] M. Pei, Y. Jia, and S.-C. Zhu, "Parsing video events with goal inference and intent prediction," in *ICCV*, 2011.
- [39] M. Hoai and F. De la Torre, "Max-margin early event detectors," in *CVPR*, 2012.

- [40] K. Kim, M. Grundmann, A. Shamir, I. Matthews, J. Hodgins, and I. Essa, "Motion fields to predict play evolution in dynamic sport scenes," in *CVPR*, 2010.
- [41] H. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," in *Robotics: Science and Systems (RSS)*, 2013.
- [42] B. Zhou, X. Wang, and X. Tang, "Random field topic model for semantic region analysis in crowded scenes from tracklets," in *CVPR*, 2011.
- [43] K. H. Lee, M. G. Choi, Q. Hong, and J. Lee, "Group behavior from video: A data-driven approach to crowd simulation," in *SCA*, 2007.
- [44] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Eurographics*, 2007.
- [45] S. Pellegrini, J. Gall, L. Sigal, and L. V. Gool, "Destination flow for crowd simulation," in *ECCV*, 2012.
- [46] D. Xie, S. Todorovic, and S. C. Zhu, "Inferring "dark matter" and "dark energy" from videos," in *ICCV*, 2013.
- [47] A. J. Brizard, *An introduction to Lagrangian mechanics*. Singapore: World Scientific, 2008.
- [48] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2281, 2012.
- [49] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *IJCV*, vol. 101, no. 1, pp. 184–204, Sep. 2013.
- [50] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *TPAMI*, vol. 24, no. 24, pp. 509–522, 2002.



**Dan Xie** received the PhD degree in Statistics from University of California, Los Angeles (UCLA) in 2016. He received his B.E. degree in Software Engineering from Zhejiang University in China in 2011. His research interests include computer vision and machine learning.



**Tianmin Shu** received his B.S. degree in electronic engineering from Fudan University, China in 2014. He is currently a Ph.D. candidate in the Department of Statistics at University of California, Los Angeles. His research interests include computer vision, robotics and computational cognitive science, especially social scene understanding and human-robot interactions. He is a student member of the IEEE.



**Sinisa Todorovic** received the PhD degree in electrical and computer engineering from the University of Florida in 2005. He is an associate professor in the School of Electrical Engineering and Computer Science at Oregon State University. He was a postdoctoral research associate in the Beckman Institute at the University of Illinois at Urbana-Champaign, between 2005 and 2008. His research interests include computer vision and machine learning problems. He is a member

of the IEEE.



**Song-Chun Zhu** received his Ph.D. degree from Harvard University. He is currently professor of Statistics and Computer Science at UCLA. His research interests include vision, statistical modeling, learning, cognition, situated dialogues, robot autonomy and AI. He received a number of honors, including the Helmholtz Test-of-time award in *ICCV* 2013, the Aggarwal prize from the *IAPR* in 2008, the David Marr Prize in 2003 with Z. Tu et al. for image parsing, twice Marr Prize honorary nominations with Y. Wu et al. in 1999 for texture modeling and 2007 for object modeling respectively. He received the Sloan Fellowship in 2001, a US NSF Career Award in 2001, and an US ONR Young Investigator Award in 2001. He is a Fellow of IEEE since 2011.