CVPR
#****

# Learning Causalities behind Human Trajectories

Kratarth Goel & Alexandre Robicquet
Paper ID ****

## Abstract

*We present a fully un-supervised framework to learn the causalities behind human navigation. Using automatically extracted trajectories from aerial videos, we learn to model the interactions between moving targets and their static surrounding. This is in contrast to traditional approaches which use handcrafted functions such as "Social forces" or focus on one type of interactions given limited labeled data.*

*We use an end-to-end trainable recurrent convolutional architecture to predict where targets move next. Thanks to both the representation power of Convolutional Neural Networks (CNN) and the Long Short-Term Memory (LSTM) recurrent network, our approach is able to infer "navigable" paths without explicitly using scene labels or recognizing the target's physical class (e.g. pedestrian, cyclist, or driver). Given the raw crop region surrounding a target, our method predicts its trajectory for the next frames. Although the used trajectories are error-prone, our learned model outperforms previous methods on public datasets as well as a newly collected one made of aerial views.*

*Prediction, forecasting, human trajectory, human-Space interaction, Long Short-Term Memory, LSTM, Convolutional Neural Network, CNN.*

## 1. Introduction

Human navigation is not random. When pedestrians and cyclists navigate their way through cities or school campuses, they respect a set of rules. They avoid each others, prefer to stay on sidewalks, and sharply turn at intersections while keeping a personal distance to their surrounding. All these behaviors obey social and safety rules. In this work, we aim to jointly learn to model these interactions between humans (referred to as *human-human*), and their static surrounding (referred to as *human-space*)[1]. The capability to model these interactions is used to predict where humans will move next. Such prediction goal is key to a wide range of applications - from the development of simulators, socially-aware robots [41], early warning systems for autonomous agents, to the design of intelligent tracking systems in smart environments [75].

[1]Note that *humans* in this work includes any moving target such as bicyclists, drivers, or skateboarders.

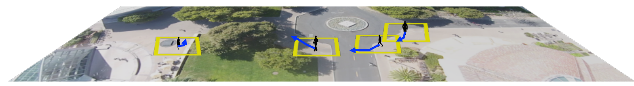

Figure 1. We propose an end-to-end trainable recurrent convolutional architecture to predict where targets move next. Given the raw crop image surrounding a target (the yellow bounding boxes), our method predicts its trajectory for the next frames (the blue arrows).

Like any other prediction task, the ability to robustly predict human navigation highly depends on the available data and the capacity of the model to jointly reason on multiple cues. To date, we argue that data has guided and constrained the design of previous methods. From Helbing *et. al.* [24], who proposed hand-crafted set of functions that mimics "social forces", to Kitani *et. al.* [32] who used static scene labels to predict the long-term trajectories, most of previous work relied on limited amount of labeled data. As a result, they accurately modeled simple interactions and were penalized to generalize to complex subtle interactions coming from mutual interactions between humans and the space.

Recently, Alahi *et. al.* [] have used ground truth pedestrians' trajectories to learn the complex interactions between humans. However, they only focus on pedestrians (as opposed to multiple classes of targets such as bicyclists, skateboards, and vehicles), and did not model the human-space interactions. Inspired by their work to develop a data-driven method to learn interactions, we propose to jointly learn both human-space and human-human interactions in a fully un-supervised framework, i.e., given error-prone trajectories.

In this work, we study the representation power of a recurrent convolutional architecture to learn to predict the long-term motion trajectory of any target. We demonstrate the performance of our method using aerial views from campus scenes where several classes of targets such as pedestrians, bicyclists, skateboarders, cars or buses interact in complex crowded environments. At training, instead of using ground truth trajectories (as previous meth-

ods), we use the output of a state-of-the-art Mutli-Target Tracking (MTT) algorithm. Thanks to the recent success of public challenges , the community has made great progress in MTT. We show that the current performance of a state-of-the-art MTT algorithm is good enough to learn to predict any target trajectory. For each tracked target, our recurrent convolutional network takes as input the surrounding raw image around the target, and outputs its future trajectory (see Figure 1).
In summary, the contributions of our paper are as follows:

(i) *Recurrent convolutional architecture for human prediction.* Inspired by the recent success of hybrid architectures that use both Convolutional Neural Networks (CNN) and Long-Short Term Memory networks (LSTM) for different sequence prediction tasks such as handwriting [] or image captioning task [20], we adapt these architectures to predict any target's motion dynamics. While LSTMs have the ability to learn and reproduce long sequences thus helping us model dependencies between multiple sequences correlated in time, the CNN with its hierarchical feature representation help us learn how this sequences interact in space, *i.e.* what is "navigable" (see Section 3).

(ii) *Un-supervised framework.* Our second contribution relies on the un-supervised nature of our learning scheme. We show that we do not need to use ground truth trajectories to learn the causalities behind navigation. Our model is robust to tracking errors.

(iii) *Campus drone dataset.* Finally, we publicly share a new dataset of UAV videos where more than 20K trajectories are labeled from 6 difference classes of targets leading to several hundred thousands of interactions. More details are available in Section 4.

In Section 5, we demonstrate the strength of our approach with respect to previous works that relies on a pre-classification step [32, **?**]. Our method is capable, from aerial views, to predict any target's trajectory without explicitly classifying its class (e.g., pedestrian, bicyclist, or car) neither its surrounding scene labels (e.g., side walk, grass, or building). It hence facilitates its usage in learning the dynamic of any other agents in other fields such as ants or mice for biological studies.

We believe that not only is this setting closer to any real world scenario, but also important for considering many interesting cases where people change their path drastically from their previous time step to accommodate the change in their static surroundings. For instance, to make a sharp right turn at a cross roads. The most interesting issue being tacked over here is that, we consider both human-human and human-space interactions at the same time, which could explain very typical cases of human behaviour where a person not only accommodate social norms and how other people are walking around them, but also the constraints imposed by their surroundings. Clearly such complex interactions cannot be modelled by hand-crafted features or heuristics. Hence we come up with a data driven approach to learn all these complex scenarios while also reasoning about subtle underlying interactions.

One of the major contribution of this paper is to introduce a hybrid model that uses in a first place moving agent detector and tracker, to extract directly from the raw images the position and trajectories of the moving agent. We then use a CNN to extract static semantics and feed it to the LSTM for jointly reasoning about the future trajectories of people in a crowded space. This architecture, which we refer to as the Space-Time Network (because it encodes information both in time and space to predict trajectories of people in the future), can automatically learn typical interactions that take place among trajectories. This model leverages existing human trajectory datasets without the need for any additional annotations to learn common sense rules and conventions that humans observe in while moving in any kind of environment.

## 2. Related work

Methods to forecast human navigation can be grouped into two categories: the ones modeling the dynamic content, *human-human* interactions, and the ones focusing on the static scene, *human-space* interactions. We briefly present an overview of past works for both approaches. We also discuss relevant Recurrent Neural Network (RNN) models for sequence prediction tasks.

**Human-human interactions.** Helbing and Molnar [24] presented a pedestrian motion model with attractive and repulsive forces referred to as the *Social Force* model. This has been shown to achieve competitive results even on modern pedestrian datasets [39, 49]. This method was later extended to robotics [41] and activitiy understanding [43, 72, 50, 38, 37, 9, 10].

Similar approaches have been used to model human-human interactions with strong priors for the model. Treuille *et. al.* [60] use continuum dynamics, Antonini *et. al.* [3] propose a Discrete Choice framework and Wang *et. al.* [67], Tay *et. al.* [59] use Gaussian processes. Such functions have alse been used to study stationary groups [73, 48]. These works target smooth motion paths and do not handle the problems associated with discretization.

Another line of work uses well-engineered features and attributes to improve tracking and forecasting. Alahi *et. al.* [1] presented a social affinity feature by learning from human trajectories in crowd their relative positions, while Yu *et. al.* [73] proposed the use of human-attributes to improve forecasting in dense crowds. They also use an agent-based model similar to [6]. Rodriguez et al. [54] analyze videos with high-density crowds to track and count people.

Most of these models provide hand-crafted energy potentials based on relative distances and rules for specific scenes. In contrast, we propose a method to learn human-human interactions in a more generic data-driven fashion.

**Human-space interactions.** Human-space models try to predict the motion and/or action to be carried out by people in a video using the static space information. A large body of work learns motion patterns through clustering trajectories [26, 30, 46, 77]. More approaches can be found in [45, 52, 34, 4, 15, 33]. Kitani *et. al.* in [32] use *Inverse Reinforcement Learning* to predict human paths in static scenes. They infer walkable paths in a scene by modeling human-space interactions. Walker in [66] predict the behavior of generic agents (*e.g.*, a vehicle) in a visual scene given a large collection of videos. Ziebart et al. [79, 23] presented a planning based approach.

Turek [61, 40] used a similar idea to identify the functional map of a scene. Other approaches like [27, 18, 42, 36] showed the use of scene semantics to predict goals and paths for human navigation. Scene semantics has also been used to predict multiple object dynamics [16, 36, 34, 28]. These works are mostly restricted to the use of static scene information to predict human motion or activity. In our work, we focus on modeling dynamic crowd interactions for path prediction.

More recent works have also attempted to predict future human actions. In particular, Ryoo *et. al.* [55, 8, 69, 65, 44, 58] forecast actions in streaming videos. More relevant to our work, is the idea of using a RNN mdoel to predict future events in videos [53, 57, 64, 56, 31]. Along similar lines, we predict future trajectories in scenes.

**CNN and LSTM models for sequence prediction.** Recently Recurrent Neural Networks (RNN) and their variants including Long Short Term Memory (LSTM) [25] and Gated Recurrent Units [12] have proven to be very successful for sequence prediction tasks. : speech recognition [20, 11, 13], machine translation [5], .At the same time both standalone Convolutional Neural Networks have also shown some success in these tasks. However where these architectures best shine is when they are part of a hybrid model that uses the advantages of both LSTMs and CNNs. image/video classification [7, 21, 68, 47], human dynamics [17] and caption generation [62, 29, 74, 14, 71] to name a few. RNN models have also proven to be effective for tasks with densely connected data such as semantic segmentation [76], scene parsing [51] and even as an alternative to Convolutional Neural Networks [63]. These works show that RNN models are capable of learning the dependencies between spatially correlated data such as image pixels. This motivates us to extend the sequence generation model from Graves et al. [19] to our setting. In particular, Graves et al. [19] predict isolated handwriting sequences; while in our work we jointly predict multiple correlated sequences corresponding to human trajectories.

## 3. Our Method - Space Time Network

The interplay between the static and dynamic content of a scene guides human navigation. For instance, a person can decide to turn because (s)he arrives at an intersection or needs to avoid a group of people moving towards him. Such deviation in trajectory cannot be predicted by observing the person's past behavior in isolation.

This motivates our work to jointly model the static surrounding of a target in addition to its dynamic one. In this section, we describe our model that uses CNN to learn a representation from the static surrounding combined with LSTM-based architecture to predict the trajectories of any target in a scene.

### 3.1. Problem formulation

We aim to predict the trajectory of a target given its observed short-term motion and surrounding visual information. Each scene is first preprocessed to obtain the spatial coordinates of the all moving targets at different time-instants using a MTT algorithm. . At any time-instant $t$, the $i^{th}$ target in the scene is represented by his/her xy-coordinates $(x_t^i, y_t^i)$ and its surrounding raw image (a crop rectangular image of 100 $m^2$ centered on the target). We observe the positions of all the targets from time 1 to $T_{obs}$, and predict their positions for time instants $T_{obs+1}$ to $T_{pred}$.

This task is similar to a sequence generation problem [19], where the input sequence corresponds to the cropped images of a target and the output sequence denotes his/her future positions at different time-instants.

### 3.2. Recurrent convolutional architecture - Space Time Network

Every target has a different motion pattern: they move with different velocities, acceleration and have different gaits. We need a model which can understand such target-specific motion properties from a limited set of initial observations.

We expect the hidden states of an LSTM to capture these time varying motion-properties, and we expect the CNN to extract rich scene semantic features that tells the LSTM how targets are interacting with the space around them. We jointly train this model to be robust towards saliency in both time and space domain.

We prove empirically that this is indeed the case and that the model is able to predict turns that are not heuristic

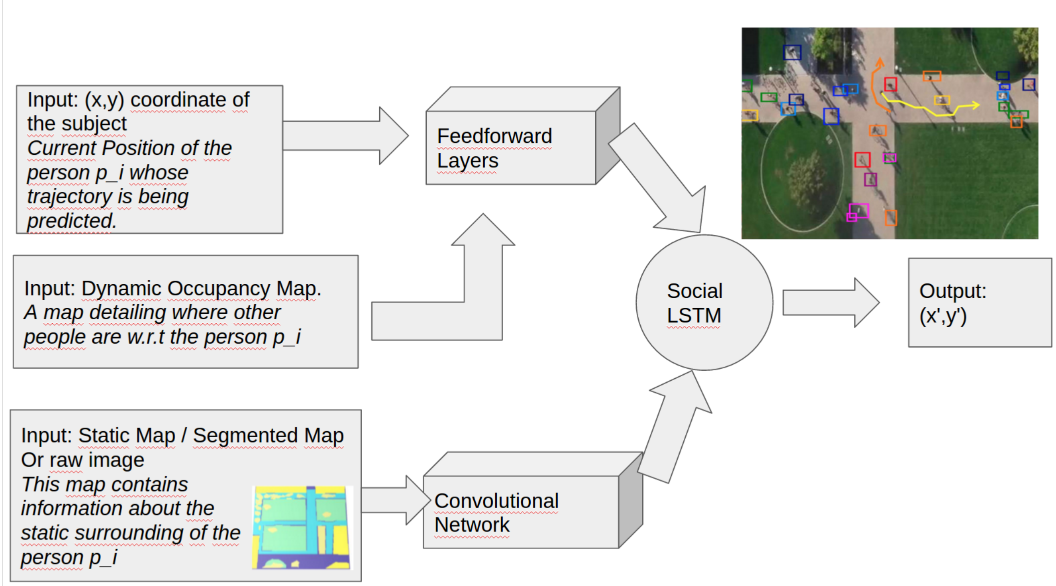


Figure 2. An overview of the hybrid CNN-LSTM model called the Space Time Network

based and are dependent on human reasoning about space and time around them.

In order to jointly reason across multiple people, we share the states between neighboring LSTMS. This introduces a new challenge: every person has a different number of neigh- bors and in very dense crowds [2], this number could be prohibitively high. Hence, we need a compact representation which combines the information from all neighboring states. We handle this by using neighborhood pooling layers. At every time-step, the LSTM cell receives pooled hidden-state information from the LSTM cells of neighbors as well as a vector representing features of the static scene around them from the CNN. While pooling the information, we try to preserve the spatial information through grid based pooling as explained below.

The hidden state $h_t^i$ of the LSTM at time $t$ captures the latent representation of the $i^{th}$ person in the scene at that instant. This representation is shared with neighbors by building a neighborhood hidden-state tensor $H_t^i$ . Given a hidden-state dimension $D$, and neighborhood size $N_o$, there is a $N_o$ $N_o$ $D$ tensor $H_t^i$ for the $i^{th}$ trajectory:

$$H_t^i(m,n,:) = \sum_{j \in N_i} \mathbf{1}_{mn}[x_t^j - x_t^i, y_t^j - y_t^i] h_{t-1}^j \qquad (1)$$

where $h_{t1}^j$ is the hidden state of the LSTM corresponding to the $j^{th}$ person at $t1$, $\mathbf{1}_{mn}[x,y]$ is an indicator function to check if $(x,y)$ is in the $(m,n)$ cell of the grid, and $N_i$ is the set of neighbors corresponding to person $i$. The pooled neighborhood hidden-state tensor is embed into a vector $a_i^t$ and the co-ordinates into $e_i^t$ .

Also we embed the scene around the person $i$ as $c_t^i$ using a convolutional network architecture. These embeddings are concatenated and used as the input to the LSTM cell of the corresponding trajectory at time $t$. This introduces the following recurrence:

$$\begin{aligned} r_t^i &= \phi(x_t^i, y_t^i; W_r) \\ e_t^i &= \phi(a_t^i, H_t^i, W_e) \\ c_t^i &= CNN(c_t^i, W_c) \\ h_t^i &= \phi(a_t^i, h_{t-1}^i, e_t^i, c_t^i; W_l) \end{aligned}$$

where $\phi(.)$ is an embedding function with ReLU non-linearity, $W_r$ and $W_e$ and $W_c$ are embedding weights. The LSTM weights are denoted by $W_l$ .

## 4. Campus Dataset

We aim to learn the remarkable human capability to navigate in complex and crowded scenes. Existing datasets mainly capture the behavior of humans in spaces occupied by a single class of object, *e.g.*, pedestrian-only scenes [49, 39, 1]. However, in practice, pedestrians share the spaces with other classes of objects such as bicyclists, or skateboarders to name a few. For instance, on university campuses, a large variety of these objects interacts at peak hours. We want to study social navigation in these complex and crowded scenes occupied by several classes of objects.

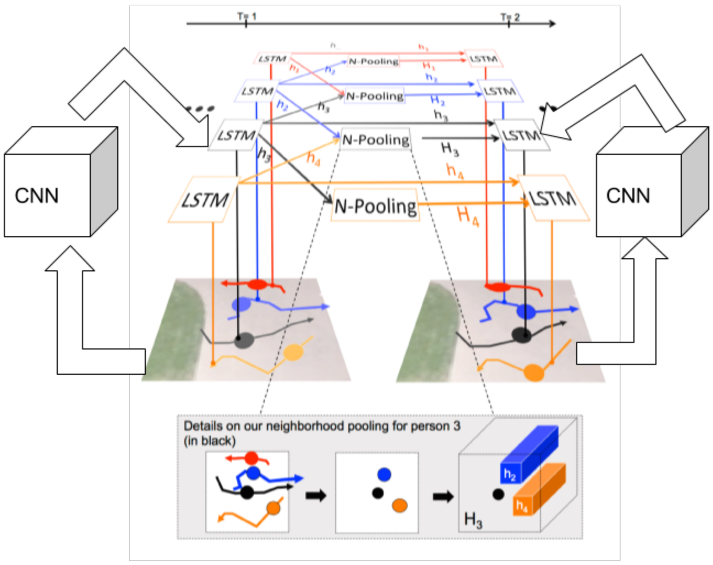


Figure 3. A more in depth representation of the Space Time Network. The Convolutional Networks are given as input a small local patch of space around the object that we want to predict the future of motion for. It is the CNNs job to extract information from these images to provide as input to the LSTM Additional the LSTM given as input the occupancy map of people around it that indicates where people are moving with respect to the persons current position and also the current (x,y) co-ordinates of the person whose future trajectory is being predicted. We use a separate LSTM network for each trajectory in a scene. The LSTMs are then connected to each other through a neighborhood pooling (N- pooling) layer. Unlike the traditional LSTM, this pooling layer allows spatially proximal LSTMs to share information with each other. The variables in the figure are explained in Eq. 2. The bot- tom row shows the N-pooling for one person in the scene. The hidden-states of all LSTMs within a certain radius are pooled to- gether and used as an input at the next time-step

To the best of our knowledge, we have collected the first large-scale dataset that has images and videos of various types of targets interacting in a real-world university campus. Our dataset captures the following types of interactions:

- target-target interactions, *e.g.*, a bicyclist avoiding a pedestrian,
- target-space interactions, *e.g.*, a skateboarder turning around a roundabout.

| Dataset | Frames | Targets | Interactions | Physical class |
|---|---|---|---|---|
| ISENGARD | 134079 | 2044 | 6472 | 6 |
| HOBBITON | 138513 | 3821 | 14084 | 6 |
| EDORAS | 47864 | 1186 | 4684 | 5 |
| MORDOR | 139364 | 4542 | 68459 | 6 |
| FANGORN | 249967 | 3126 | 45520 | 6 |
| THE VALLEY | 219712 | 4845 | 46062 | 6 |
| TOTAL | 929499 | 19564 | 185281 | 6 |

Table 1. Our campus dataset characteristics. We group the scenes and refer to them using fictional places from the "Lord of the Rings".

**Target-target interactions** We say that two targets interact when their collision energy (described by [49]) is nonzero, *e.g.*, a pedestrian avoiding a skateboarder. These interactions involve multiple physical classes of targets (pedestrians, bicyclists, or skateboarders to name a few), resulting into 185K annotated target-target interactions. We intentionally collected data at peak hours (between class breaks in our case) to observe high density crowds. For instance, during a period of 20 seconds, we observe in average from 20 to 60 targets in a scene (of approximately $900m^2$).

**Target-space interactions.** We say that a target interacts with the space when its trajectory deviates from a linear one in the absence of other targets in its surrounding, *e.g.*, a skateboarder turning around a roundabout. To further analyze these interactions, we also labeled the scene semantics of more than 100 static scenes with the following labels: road, roundabout, sidewalk, grass, building, and bike rack (see Figure 4). We have approximately 40k "target-space" interactions.

To the best of our knowledge, it is the first dataset to depict complex interactions at such a scale. Tables 1 and 2 present more details on our collected dataset. The scenes are grouped into 6 areas based on their physical proximity on campus. The dataset comprises more than 19K targets consisting of 11.2K pedestrians, 6.4K bicyclists, 1.3k cars, 0.3K skateboarders, 0.2K golf carts, and 0.1K buses.

Each scene is captured with a 4k camera mounted on a quadrotor platform hovering above various intersections on a University campus at an altitude of approximately eighty meters. The videos are also available for further research in detection, recognition, tracking from UAV data. The videos have been processed (*i.e.* undistorted and stabilized), and annotated with their class label and their trajectory in time and space is identified.

Our dataset can be used to conduct research in activity and scene understanding. For example, the collected trajectories can be used to infer the functionality map of a scene [22, 70, 78, 35], *e.g.*, infer sitting areas, and improve image segmentation. We envision our dataset to be an ideal testbed for pushing the limits of visually intelligent machines. It enables the design of new methods that allow learning multi-target interactions at a large scale as well as pushing research on multi-target tracking.

| Dataset | Bi | Ped | Skate | Carts | Car | Bus |
|---|---|---|---|---|---|---|
| ISENGARD | 1004 | 926 | 57 | 19 | 23 | 15 |
| HOBBITON | 163 | 2493 | 24 | 18 | 1065 | 58 |
| EDORAS | 224 | 956 | 2 | 2 | 2 | 0 |
| MORDOR | 2594 | 1492 | 111 | 154 | 165 | 26 |
| FANGORN | 1017 | 1991 | 50 | 30 | 27 | 11 |
| THE VALLEY | 1362 | 3358 | 89 | 21 | 10 | 5 |
| TOTAL | 6364 | 11216 | 333 | 244 | 1292 | 115 |

Table 2. Details on the number of objects in our campus dataset. Bi = bicyclist, Ped = pedestrian, Skate = skateboarders.

# 5. Results and Experiments

## 5.1. Training

We train the our architecture the Space-Time Network on three different settings. We use two different kinds of trajectories. 1) we produce synthetic trajectories by simulating an number of pedestrians, using a Social Force model. We made this model "multi-class" by using 3 different sets of Social Parameters. 2) We use real world trajectories captured from the campus dataset. Also we use two different type of bakcgrounds. 1) black and white , walkable and non walkable, image patches prepared from intuitive understanding of scene semantics, i.e. cross roads, bridges etc, and trajectories overlayed on the map based on social-force principles. 2) Real world images of places from the campus dataset that are pre-segmented using .

Thus we experiment on three datasets whose description are as follows:- 1) Data1: black and white (walkable , non-walkable) regions with synthetic trajectories. 2) Data2: Pre-Segmented static scene map from campus dataset with synthetic trajectories. 3) Data3: Pre-Segmented static scene map from campus dataset with real world trajectories also from the campus dataset. The following is summarization of the experiments and the results we got:

Some of the models in the above table are detailed below. A baseline-LSTM is a vanilla LSTM model with 256 hidden units and tanh activations. The 'social' LSTM model is the LSTM model that associates a LSTM cell to each of pedestrain walking in the scene. This LSTM model does not take as input the static segmented scene map as input. Each of these LSTM communicates with the other LSTM by way of sharing weights and also by the 'neighborhood' pooling layer that is explained above. These LSTMs also use 256 hidden units in its cell. The Feedforward Space-Time Network uses a feedforward layer instead of the convolutional neural netowrk to extract static scene semantics. The segemented scene is passed through 3 feedforward ReLU layers (4096,1025,256) to extract semantic information about the human-space interactions. The best performing model is the Space-Time Network model that replaces the feedforward layers in the above Feedforward social LSTM model with a 6 layeres Convolutional Neural Network with the following architectural details. (Conv-BN-Relu-Pool)*2 — (Conv-Relu-Pool) — Feedforward*3.

# 6. Conclusion

We have presented a hybrid model called Space-Time Network that can jointly reason across multiple individuals to predict human trajectories in a scene. We use one LSTM for each trajectory and share the information between the LSTMs through the a neighborhood pooling layer. We also extract static scene semantics using a convolutional neural network. Our proposed model outperforms state-of-the-art methods on publicly available datasets. In addition, we qualitatively show that our model successfully predicts various non-linear behaviors arising from social interactions as well as human-space interactions Future work will extend our model to multi-class settings where several objects such as bicycles, skateboards, carts, and pedestrians share the same space. Each object will have its own label in the occupancy map.

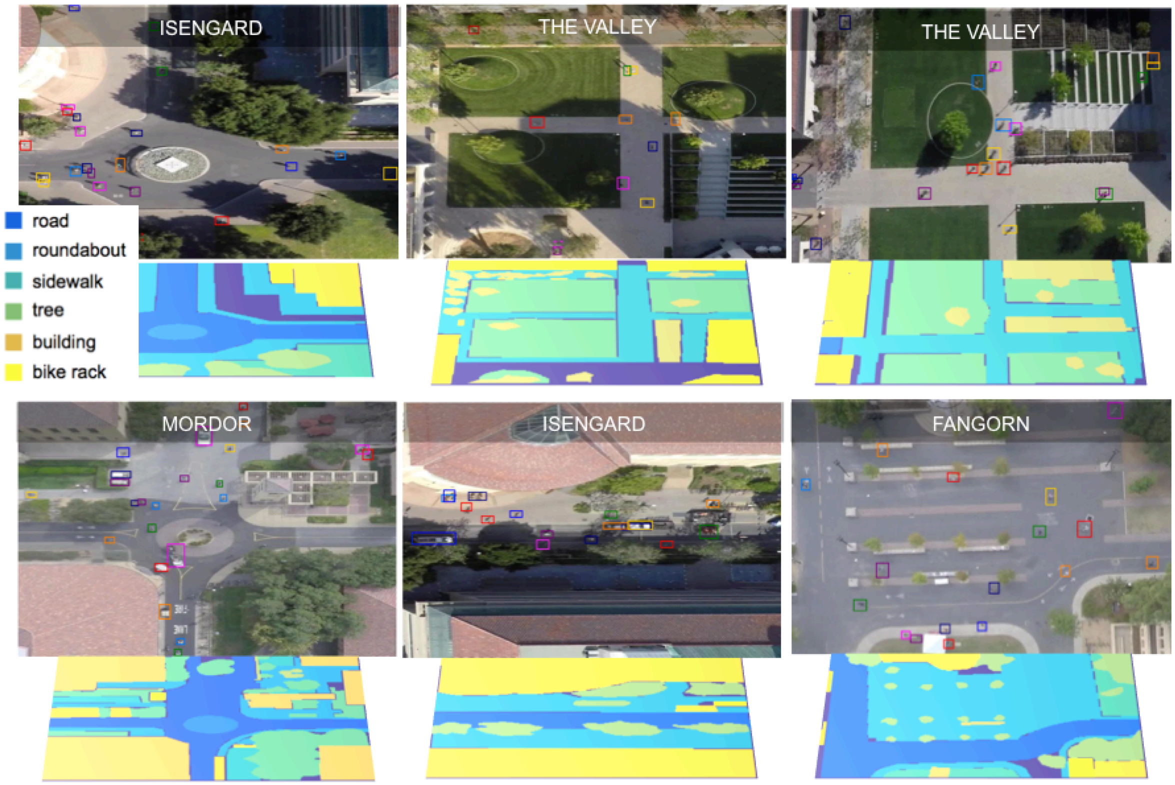


Figure 4. Some examples of the scenes captured in our dataset. We have annotated all the targets (with bounding boxes) as well as the static scene semantics (rows 2, 4, and 6). The color codes associated to target bounding boxes represents different track IDs.

| Experiment | Dataset | | | | | |
|---|---|---|---|---|---|---|
| | Data1 | | Data2 | | Data3 | |
| | NLL | AD Err | NLL | AD Err | NLL | AD Err |
| Baseline LSTM | 3.3614 | 5.3259 | 5.5124 | 7.8951 | 6.2158 | 8.9657 |
| Social-LSTM w/o static scene map | 1.9524 | 2.1985 | 3.5548 | 4.0370 | 3.7804 | 4.9935 |
| Space-Time Network (feedforward) | -2.1578 | 0.9882 | -1.2208 | 1.2015 | -1.0632 | 1.4429 |
| Space-Time Network (CNN) | -10.215 | 0.5547 | -8.8521 | 0.6215 | -8.6004 | 0.2854 |

Table 3. Quantitative Results (AD-Err holds for the Average Displacement error in meters between the predicted trajectory and the ground truth for synthetic data )

## References

[1] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014.

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2211–2218. IEEE, 2014.

[3] G. Antonini, M. Bierlaire, and M. Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006.

[4] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, and A. Oliver-Albert. A predictive model for recognizing human behaviour based on trajectory representation. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 1494–1501. IEEE, 2014.

[5] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] E. Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7280–7287, 2002.

[7] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, et al. Look and


Figure 5.

think twice: Capturing top-down visual attention with feedback convolutional neural networks. *ICCV*, 2015.

[8] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. M. Siskind, and S. Wang. Recognize human activities from partially observed videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2658–2665. IEEE, 2013.

[9] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Computer Vision–ECCV 2012*, pages 215–230. Springer, 2012.

[10] W. Choi and S. Savarese. Understanding collective activitiesof people from videos. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(6):1242–1257, 2014.

[11] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv preprint arXiv:1412.1602*, 2014.

[12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[13] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. *CoRR*, abs/1506.02216, 2015.

[14] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*, 2014.

[15] J. Elfring, R. Van De Molengraft, and M. Steinbuch. Learning intentions for improved human motion prediction. *Robotics and Autonomous Systems*, 62(4):591–602, 2014.

[16] D. F. Fouhey and C. L. Zitnick. Predicting object dynamics in scenes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2027–2034. IEEE, 2014.

[17] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics.

[18] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 619–626, Washington, DC, USA, 2011. IEEE Computer Society.

[19] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[20] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1764–1772, 2014.

[21] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

[22] A. Gupta, A. Kembhavi, and L. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(10):1775–1789, 2009.

[23] K. P. Hawkins, N. Vo, S. Bansal, and A. F. Bobick. Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration. In *Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on*, pages 499–506. IEEE, 2013.

[24] D. Helbing and P. Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

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

[26] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank. Semantic-based surveillance video retrieval. *Image Processing, IEEE Transactions on*, 16(4):1168–1181, 2007.

[27] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008.

[28] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In *Computer Vision–ECCV 2014*, pages 489–504. Springer, 2014.

[29] A. Karpathy et al. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014.

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

[31] K. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3241–3248, June 2011.

[32] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert. Activity forecasting. *Computer Vision–ECCV 2012*, pages 201–214, 2012.

[33] Y. Kong, D. Kit, and Y. Fu. A discriminative model with multiple temporal scales for action prediction. In *Computer Vision–ECCV 2014*, pages 596–611. Springer, 2014.

[34] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila. Context-based pedestrian path prediction. In *Computer Vision–ECCV 2014*, pages 618–633. Springer, 2014.

[35] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Trans. Pattern Anal. Mach. Intell.*, in press, 2015.

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

[37] L. Leal-Taixé, M. Fenzi, A. Kuznetsova, B. Rosenhahn, and S. Savarese. Learning an image-based motion context for multiple people tracking. In *CVPR*, pages 3542–3549. IEEE, 2014.

[38] L. Leal-Taixé, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 120–127. IEEE, 2011.

[39] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[40] K. Li and Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(8):1644–1657, 2014.

[41] M. Luber, J. A. Stork, G. D. Tipaldi, and K. O. Arras. People tracking with human motion predictions from social forces. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 464–469. IEEE, 2010.

[42] D. Makris and T. Ellis. Learning semantic scene models from observing activity in visual surveillance. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 35(3):397–408, 2005.

[43] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE, 2009.

[44] B. Minor, J. R. Doppa, and D. J. Cook. Data-driven activity prediction: Algorithms, evaluation methodology, and applications. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 805–814. ACM, 2015.

[45] B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(8):1114–1127, 2008.

[46] B. T. Morris and M. M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2287–2301, 2011.

[47] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. *arXiv preprint arXiv:1503.08909*, 2015.

[48] H. S. Park and J. Shi. Social saliency prediction.

[49] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 261–268. IEEE, 2009.

[50] S. Pellegrini, A. Ess, and L. Van Gool. Improving data association by joint modeling of pedestrian trajectories and groupings. In *Computer Vision–ECCV 2010*, pages 452–465. Springer, 2010.

[51] P. H. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene parsing. *arXiv preprint arXiv:1306.2795*, 2013.

[52] H. Pirsiavash, C. Vondrick, and A. Torralba. Inferring the why in images. *arXiv preprint arXiv:1406.5472*, 2014.

[53] M. Ranzato et al. Video (language) modeling: a baseline for generative models of natural videos. *arXiv:1412.6604*, 2014.

[54] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert. Data-driven crowd analysis in videos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1235–1242. IEEE, 2011.

[55] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1036–1043. IEEE, 2011.

[56] M. Ryoo, T. J. Fuchs, L. Xia, J. Aggarwal, and L. Matthies. Early recognition of human activities from first-person videos using onset representations. *arXiv preprint arXiv:1406.5309*, 2014.

[57] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. *arXiv:1502.04681*, 2015.

[58] A. Surana and K. Srivastava. Bayesian nonparametric inverse reinforcement learning for switched markov decision processes. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*, pages 47–54. IEEE, 2014.

[59] M. K. C. Tay and C. Laugier. Modelling smooth paths using gaussian processes. In *Field and Service Robotics*, pages 381–390. Springer, 2008.

[60] A. Treuille, S. Cooper, and Z. Popović. Continuum crowds. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 1160–1168. ACM, 2006.

[61] M. W. Turek, A. Hoogs, and R. Collins. Unsupervised learning of functional categories in video scenes. In *ECCV*, 2010.

[62] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[63] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio. Renet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*, 2015.

[64] C. Vondrick, H. Pirsiavash, and A. Torralba. Anticipating the future by watching unlabeled video. *arXiv preprint arXiv:1504.08023*, 2015.

[65] T.-H. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. Predicting actions from static scenes. In *Computer Vision–ECCV 2014*, pages 421–436. Springer, 2014.

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

[67] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(2):283–298, 2008.

[68] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. *arXiv preprint arXiv:1411.6447*, 2014.

[69] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring" dark matter" and" dark energy" from videos. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2224–2231. IEEE, 2013.

[70] D. Xie, S. Todorovic, and S.-C. Zhu. Inferring "dark matter" and "dark energy" from videos. In *ICCV*, 2013.

[71] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015.

[72] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1345–1352. IEEE, 2011.

[73] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3488–3496, 2015.

[74] D. Yoo, S. Park, J.-Y. Lee, A. Paek, and I. S. Kweon. Attentionnet: Aggregating weak directions for accurate object detection. *arXiv preprint arXiv:1506.07704*, 2015.

[75] X. Zhang, S. Yang, Y. Y. Tang, and W. Zhang. Crowd motion monitoring with thermodynamics-inspired feature. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[76] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. *arXiv preprint arXiv:1502.03240*, 2015.

[77] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3441–3448. IEEE, 2011.

[78] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014.

[79] 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 *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 3931–3936. IEEE, 2009.