Forecasting People Trajectories and Head Poses by Jointly Reasoning on Tracklets and Vislets

Irtiza Hasan, Francesco Setti, Theodore Tsesmelis, Vasileios Belagiannis, Sikandar Amin, Alessio Del Bue, Marco Cristani, and Fabio Galasso

Abstract— In this work, we explore the correlation between people trajectories and their head orientations. We argue that people trajectory and head pose forecasting can be modelled as a joint problem. Recent approaches on trajectory forecasting leverage short-term trajectories (*aka* tracklets) of pedestrians to predict their future paths. In addition, sociological cues, such as expected destination or pedestrian interaction, are often combined with tracklets. In this paper, we propose MiXing-LSTM (MX-LSTM) to capture the interplay between positions and head orientations (vislets) thanks to a joint unconstrained optimization of full covariance matrices during the LSTM backpropagation. We additionally exploit the head orientations as a proxy for the visual attention, when modeling social interactions. MX-LSTM predicts future pedestrians location and head pose, increasing the standard capabilities of the current approaches on long-term trajectory forecasting. Compared to the state-of-the-art, our approach shows better performances on an extensive set of public benchmarks. MX-LSTM is particularly effective when people move slowly, *i.e.* the most challenging scenario for all other models. The proposed approach also allows for accurate predictions on a longer time horizon.

Index Terms—LSTM, Trajectory Forecasting, RNN, head pose estimation, visual attention, gaze estimation.

1 Introduction

Pedestrian forecasting stands for anticipating the future, based on observations and on prior understanding of the scene and actors. Further to past trajectories, forecasting the position of pedestrians requires therefore an intuition of the people goals [60], their social interaction models [3], [31], [67], the understanding of their behavior [6], [51], [55] and possible interactions with the scene [46].

Forecasting is important for tracking [54], [69], [84], especially in the case of missing or sparse target observations. In addition, it is a crucial compound for early action recognition [40], [68], [82] and more in general for surveillance systems [17], [23]. Furthermore it is indispensable for deploying autonomous vehicles, which should avoid collisions [11], and for conceiving robots, respectful of the human proxemics [21], [33], [47], [56], [75], [87].

Forecasting trajectories from images, however, is a complex problem and, probably for this reason, it has only recently emerged as a popular computer vision research topic. In particular, the modern re-visitation of Long Short Term Memory (LSTM) architectures [41], has enabled a leap forward in performance [31], [34], [72], [73], [78]. On one side, LSTM has allowed a seamless encoding of the social interplay among pedestrians [3], [31]. On the other side, the new systems have abandoned cues demanding *oracle* knowledge, such as the person destination point [60], and are therefore causal predictions.

In this paper, we differ from previous approaches, because we additionally leverage the visual attention of people for forecasting, further to their position. We infer their visual attention from their head pose. We are motivated by the strong correlation between the past short-term trajectories of the people (sequences of (x, y) position coordinates, named *tracklets*) and their corresponding sequences of head pan orientations, which we name *vislets*. Our novel contribution is supported by several sociological studies [13], [19], [24], [25], [26], [59], [77]and here motivated by statistical analysis conducted on the UCY dataset [52], which we report in Section 2.

This work introduces MiXing LSTM (MX-LSTM), an LSTM-based framework that encodes the relation between the movement of the head and people dynamics. For example, it captures the fact that rotating the head towards a particular direction may anticipate turning and starting to walk (as in the case of a person leaving a group after a conversation). This is achieved in MX-LSTM by mixing the tracklet and vislet streams in the LSTM hidden state recursion by means of a cross-stream full covariance matrix. During the LSTM backpropagation, the covariance matrix is constrained to be positive-semidefinite by means of a log-Cholesky parameterization. This model generalizes the approach of [3] (specific to the 2D positions *x,y* of people) to model state variables of dimensions four (position and head pose) and higher.

Vislets allow for a more informative social interplay among people. Instead of considering all pedestrians within a radius, as done in [3], [78], here we only consider those individuals whom the person can see. Furthermore MX-LSTM forecasts both tracklets and vislets. Predicting visual attention in crowded scenarios makes a novel frontier for research and new applications.

We have first presented MX-LSTM in [36]. This paper extends our previous work in four directions: 1) we include a comprehensive evaluation of its performance on the UCY video

- *I. Hasan, F. Setti, and M. Cristani are with the Department of Computer Science, University of Verona, Verona, Italy. E-mail: irtiza.hasan@univr.it*
- *T. Tsesmelis and A. Del Bue are with the Italian Institute of Technology, Genova, Italy.*
- *I. Hasan, T. Tsesmelis, S. Amin, and F. Galasso are with Osram GmbH, Munich, Germany.*
- *V. Belagiannis, is with Ulm University, Ulm, Germany. This work was done while working at Osram GmbH.*

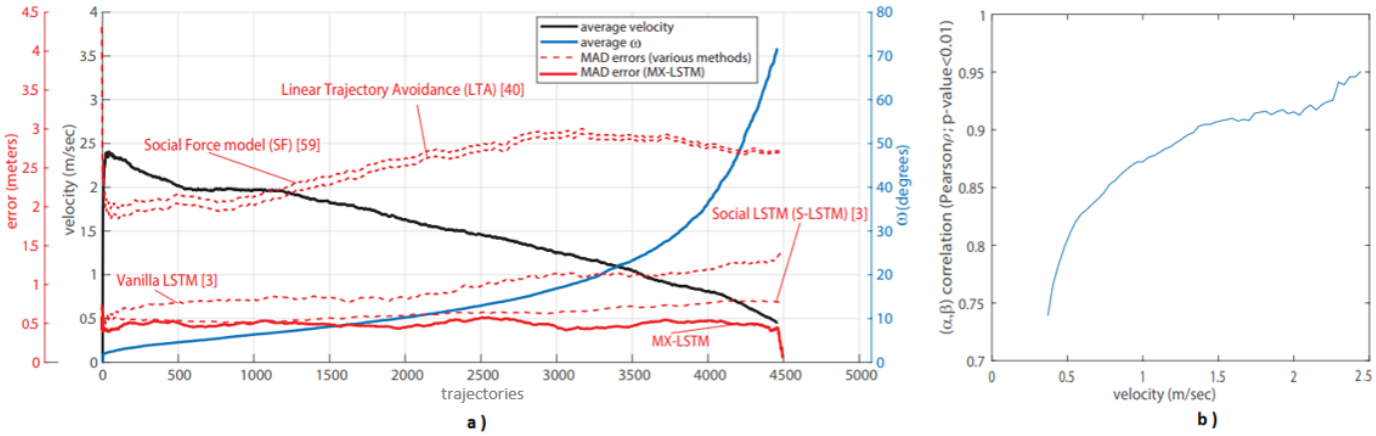

Fig. 1. Motivating the MX-LSTM: a) analysis between the angle discrepancy ω between head pose and movement, the pedestrain velocity and the average errors of different approaches on the UCY sequence [52]; b) correlation between movement angle β and head orientation angle α when the velocity is varying (better in color).

sequences (Zara01, Zara02 and UCY) [52] and on the TownCentre dataset [9], following standard evaluation protocols of trajectory forecasting [3], [31], [60]. 2) we provide an extensive evaluation with the most recent approaches to show that MX-LSTM retains overall the best performance. MX-LSTM has the ability to forecast people when they are moving slowly, the Achilles heel of all the other approaches proposed so far. Additionally, here we provide novel experiments to test its robustness by predicting in the longer-term horizon and by using an estimated (thus noisy) head pose estimator [49]. In particular we quantify the performance of head pose estimates *vs.* manual labels both at training and inference. 3) We verify that vislets help beyond the mere larger model capacity, by testing MX-LSTM with position-related variables replacing vislets. 4) we provide novel qualitative illustrations, detail failure cases; and finally we perform novel simulations, which uncover how the learned head poses affect the people motion.

The rest of the paper is organized as follows: In Section 2, we motivate the need for MX-LSTM showing the correlation between head pose and trajectories in the most popular forecasting datasets. Section 3 presents the related literature. In Section 4, we present our MX-LSTM approach. Section 5 illustrates quantitative and qualitative experiments and ablation studies. Finally, Section 6 concludes the paper.

2 Motivation for the MX-LSTM

Intuitively, the head pose of person is a cue for the direction in which she/he moves. However, the literature in trajectory forecasting lacks a quantitative study on the importance of the head pose. Here we examine the common forecasting datasets to study the relationship between the head pose and motion directions. In particular, we focus on the UCY dataset [52], composed by the Zara01, Zara02 and UCY sequences, which provides the annotations for the pan angle of the head pose of all the pedestrians. We also consider the Town Center dataset [10], where we have manually annotated the head pose, using the same annotation protocol as in [52].

In this section, with specific reference to Figure 1, we present the preliminary analysis and observations, which have motivated the design of our MX-LSTM. We would specifically refer to the UCY video sequence (but similar observations applied to all others).

1) People watch their steps. We show this fact by plotting in Fig. 1a the angular discrepancy ω (*blue curve*), between the head pose α and the person motion angle β, against the velocity (*black curve*), intended as the modulus of the motion vector $\overrightarrow{\boldsymbol{x_{t+1}} - \boldsymbol{x_t}}$.

In more details, we have computed the average angular discrepancy ω for each of the people trajectories of the UCY video sequence (for each trajectory, we average ω across all frames where it occurs). In Fig. 1a, we have then arranged the trajectories in ascending order (the x axis) according to their average discrepancy angle ω (the *blue* y-axis on the sub-figure right side, marked as "ω". Please refer to Fig. 2c, for pictorial illustration of "ω"). For each trajectory we have then plotted the corresponding average speed (*black curve*), as measured on the black y-axis marked as "velocity"[1].

As it shows from Fig. 1a, 75% of the people only turn their head by 20°. They watch therefore their steps, especially at higher speeds.

2) Head pose and movements are (statistically) correlated. On Fig. 1a, we report the velocity curve (black solid line and axis). To plot this curve, we order all the trajectories with respect to the average speed of each individual. First of all, notice that the ω and the pedestrian speed are inversely proportional: the alignment between the head pose and the direction of movement is higher when the speed is higher; when the person slows down the head pose is dramatically misaligned. Secondly, the relation is statistically significant: we consider the Pearson circular correlation coefficient [44] between the angles α_t and β_t. Overall, the correlation is 0.83 (p-value$<$ 0.01), computed for all the frames of the sequences considered for Fig. 1. The plot in Fig. 1b elaborates that the correlation is lower at low velocities, where the discrepancy between the α_t and β_t angles is typically higher.

One of the challenges here, is to investigate whether the dynamic discrepancy between the head pose angle α_t and movement direction β_t at different speeds of the human motion can be learned by our proposed MX-LSTM to improve the forecasting. Moreover, MX-LSTM should learn how these relations evolve in time, which has not been investigated yet. In fact, prior work has only addressed single frames.

1. We disregard those frames where the average speed of person movement is below 0.45m/sec, since those people do not essentially move and their motion angle β can hardly be determined.

3) Forecasting is difficult for pedestrians at low speeds. In Fig. 1a (red lines and red axis), we compare the Mean Average Displacement (MAD) error [60] of the following approaches: SF [84], LTA [75], vanilla LSTM and Social LSTM [3], against our proposed MX-LSTM approach (solid red curve). We notice that lower velocities correspond generally to higher forecasting errors. When people walk slowly, their behavior becomes less predictable, not only due to physical reasons (less inertia), but also behavioral (people walking slowly are usually involved in secondary activities, such as looking around or chatting with others). By contrast, our proposed approach MX-LSTM (solid red curve) performs well even at lower velocities, since it makes use of the evidence from the head pose. MX-LSTM approaches an error close to zero for the nearly static people, as it should ideally be (more details in Sec. 5).

Summarizing, the head pose is correlated with the movement. When people move fast, this correlation is stronger and their head pose is largely aligned with the direction of motion. However, when people move slowly, the correlation is weaker (but still significant), and the head pose is drastically misaligned with the movement. This results in higher prediction errors for most state-of-the-art approaches of trajectory forecasting. These facts justify and motivate our objective with the MX-LSTM, to capture the head pose information jointly with the movement and use it for a better and more uniform trajectory forecasting, for people moving at both lower or higher speeds.

3 Related work

Trajectory forecasting [8], [58] has been traditionally addressed by approaches such as Kalman filter [45], linear [57] or Gaussian regression models [64], [65], [80], [81], auto-regressive models [2] and time-series analysis [63]. The main limitation of these approaches is the lack of modelling the human-human interactions [5], [15], [16], [48], [76], that instead plays an important role. More recent approaches have proposed to use convolutional neural networks [42], generative models [32] and recurrent neural networks [3] for modelling the trajectory prediction, which also consider the human-human interaction. We discuss these most recent related approaches in the respective subsections below.

Human-human interactions. Helbing and Molnar [38] have considered for the first time the effect of other pedestrians to the behavior of an individual. The pioneering idea has been further developed by [52], [55] and [60], who have respectively introduced a data-driven, continuous and game theoretical model. Notably, these approaches successfully employed the essential cues for track prediction, such as the human-human interaction and people intended destination. More recent works encode the human-human interactions into a "social" descriptor [4] or propose human attributes [85] for the forecasting in crowds. Other related methods [3], [78] embed the proxemic reasoning into an LSTM-based predictor. Here the social aspect is implicitly addressed by pooling the hidden LSTM variables of all actors participating in the motion. Our work mainly differentiates from [3], [52], [60], [78] because we only consider for interactions those people who are within the cone of attention of the person, (as also verified by psychological studies [43]).

Destination-focused path forecast. Path forecasting has also been framed as an inverse optimal control (IOC) problem by Kitani *et. al.* [46]. The follow-up works [1], [86] have adopted inverse reinforcement learning and dynamic reward functions [51] to address the occurring changes in the environment. We describe these approaches as destination-focused, because they require the end-point of the person track to be known. To eliminate this constraint, similar works have relaxed the destination end-point to a set of plausible path ends [21], [56]. By contrast, our approach does not require this information and it is therefore causal (while knowledge of end-point would require knowing the future).

Head pose as social motivation. Our interest into the head pose stems from sociological studies such as [13], [19], [24], [25], [26], [59], [77], whereby the head pose has been shown to correlate to the person destination and pathway. Interestingly, the correlation is higher in the cases of poor visibility, such as at night time, and in general when the person is being busy with a secondary task (*e.g.* bump avoidance) further to the basic walking [24], [25]. In our experimental studies, we observe that the head pose is correlated with the movement, especially at high velocities, while slowing down this correlation decreases too, but still remains statistically significant. These studies motivate the use of the head pose as proxy to the track forecasting.

There is prior work on estimating the head pose of people in real-time, applicable to people at low resolution [7], [27], [37], [49], [66], [71], [74]. We leverage these methods within MX-LSTM, to gather the required head pose information from the input frames (just). To the best of our knowledge, there is no prior work using head pose to forecast the pedestrian trajectories, further to our own. In [35], we integrate the view frustum of attention into an objective energy formulation. By contrast, the proposed LSTM-based framework provides an implicit data-driven joint formulation, which outperforms our previous method. In [36], we introduce MX-LSTM for the first time. Here we extend it with novel quantitative and qualitative evidence.

LSTM models. LSTM models [41] have been employed in tasks where the output is conditioned on a varying number of inputs [30], [79], notably hand writing generation [29], tracking [18], action recognition [22], [53], future prediction [42], [50], [70] and path prediction [83].

As for trajectory forecasting, Alahi *et. al.* [3] model the pedestrians as LSTMs that share their hidden states through a "social" pooling layer, avoiding to forecast colliding trajectories. This idea has been successfully adopted by [78]. In [69], it has been extended for modeling the tracking dynamics. A similar approach [34], [72] has been embedded directly in the LSTM memory unit as a regularization, which models the local spatio-temporal dependency between neighboring pedestrians. In this work, we propose a variant of the social pooling by considering a visibility attention area, defined by the head pose.

In most cases, the training of forecasting LSTMs is driven by the minimization of negative log-likelihoods whereby the probabilities are Gaussians [3], [78] or mixture of Guassians [29]. In general, when it comes to Gaussian parameters, only bidimensional data (i.e. (x, y) coordinates) have been considered so far, leading to the estimation of 2 x 2 covariance matrices. These can be optimized without considering the positive semidefinite requirement [28], that is one of the most important problems for the covariances obtained by optimization [61] (see Sec. 4.4). Here, we study the problem of optimizing Gaussian parameters of higher dimensionality for the first time.

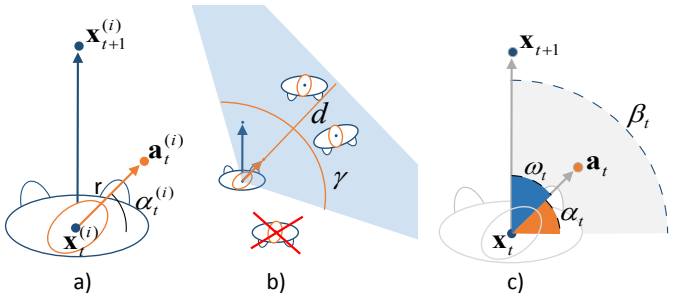

Fig. 2. A graphical interpretation of tracklets and viselets. a) tracklets $\mathbf{x}_t^{(i)}$ and $\mathbf{x}_{t+1}^{(i)}$ and vislet anchor point $\mathbf{a}_t^{(i)}$; b) Social pooling leveraging the Visual Frustum of Attention; c) angles for the correlation analysis.

4 Our Approach

In this section, we present *MX-LSTM*. The model may jointly forecast individuals' locations and pose by leveraging the information about the recent history of head positions (*tracklets*) and orientations (*vislets*). We first define the concepts of tracklets and vislets (Sec. 4.1); then we describe our proposed formulation of social pooling based on visual frustum of attention (Sec. 4.2); finally, we report details about the LSTM formulation (Sec. 4.3) and model training by optimizing the multidimensional co-variance matrices (Sec. 4.4).

4.1 Tracklets and vislets

We define as *tracklet* the list of consecutive locations on the ground plane visited by an individual during the last time steps. Formally, the tracklet associated with the i-th subject at time T is $\{\mathbf{x}_t^{(i)}\}_{t=1,\dots,T}$, where $\mathbf{x}_t^{(i)} = (x, y) \in \mathcal{R}^2$. Similarly, a *vislet* is the list of anchor points located at a fixed distance r from the subject, aligned with its head head orientation. Thus, for subject i at time T the vislet is $\{\mathbf{a}_t^{(i)}\}_{t=1,\dots,T}$, with $\mathbf{a}_t^{(i)} = (x_t^{(i)} + \cos\alpha_t^{(i)}, y_t^{(i)} + \sin\alpha_t^{(i)}) \in \mathcal{R}^2$ (see Fig. 2a).

In theory, one could encode the head orientation by means of the pan angle at each time step. We prefer to use anchor points instead, which gives several benefits. The main advantage of using vislets instead of encoding the head orientation directly with the pan angle, is that this formulation implicitly solve all the issues generated by the discontinuity between 360° and 0°. Moreover, vislets and tracklets have very similar representations, which is very convenient for modeling the interplay of these two components in the MX-LSTM structure. Please note that the distance r is irrelevant, as long as it is a constant value; in this work we set it at 0.5m for the sake of visualization.

Our method relies on a set of location and head pose observations to predict tracklets and vislets for the following estimation period. In particular, MX-LSTM mixes together the two streams to understand their relationship, providing a joint prediction. Accordingly to the trajectory forecasting literature [3], [75], [84], we consider these observations as provided by an oracle, *i.e.* given by an annotator. To directly compare our approach with the other recent ones, we provide experiments where the past head poses are estimated by a real "static" head pose estimator; in this way, MX-LSTM will require no additional effort in annotation with respect to former approaches.

We instantiate an LSTM model for each individual by using two separate embedding functions for tracklets (1) and vislets (2):

$$\mathbf{e}_t^{(x,i)} = \phi\left(\mathbf{x}_t^{(i)}, \mathbf{W}_x\right) \quad (1)$$

$$\mathbf{e}_t^{(a,i)} = \phi\left(\mathbf{a}_t^{(i)}, \mathbf{W}_a\right) \quad (2)$$

where the embedding function ϕ is the linear projection, via the embedding weigths $\mathbf{W}_{(.)}$, into a D-dimensional vector, with D the dimension of the hidden space. This is followed by a ReLU activation function.

4.2 VFoA social pooling

The concept of social pooling was first introduced by [3] as an effective way to capture (and embed into an LSTM model) how people move in a crowded space to avoid collisions. In its original form, it is an isotropic area of interest surrounding the target individual. The LSTM hidden variables of the people within the area of interest are pooled, *i.e.* collected to account for the human-human interaction. This formulation implicitly assumes that a person's trajectory is affected not only by the behaviour of people walking in front of him/her, but also by people behind him/her back as also illustrated in Fig. 3. In this paper we upgrade this model by exploiting vislet information, building on the concept of View Frustum of Attention (VFoA), that is a region where the attention of a person is focused, according to its gaze direction. We propose to model the VFoA as a circular sector originating in the head position ($\mathbf{x}_t^{(i)}$), aligned with the head pose (*i.e.* towards the anchor point $\mathbf{a}_t^{(i)}$), with a aperture angle γ; to account for the limitations of human vision in focusing on very far ahead objects, we limit the region with a maximum distance d. We learned both γ and d parameters at training time by cross-validation on the training partition of the TownCentre dataset. A graphical interpretation of the VFoA is provided in Fig. 2(b).

Formally, we define an area of interest as the squared region centered at the pedestrian location with size $2d \times 2d$; this area is then divided in a uniform grid of $N_o \times N_o$ cells. Our VFoA social pooling is a $N_o \times N_o \times D$ tensor $\mathbf{H}$ defined as follows:

$$\mathbf{H}_t^{(i)}(m, n, :) = \sum_{j \in \text{VFoA}_i} \mathbf{h}_{t-1}^{(j)}, \quad (3)$$

where the m and n indices run over the $N_o \times N_o$ grid and the condition $j \in \text{VFoA}_i$ is satisfied when the subject j is in the VFoA of subject i, $\mathbf{h}$ is the hidden state of the LSTM model. The pooling vector is then embedded into a D-dimensional vector by

$$\mathbf{e}_t^{(H,i)} = \phi(\mathbf{H}_t^{(i)}, \mathbf{W}_H). \quad (4)$$

4.3 LSTM recursion

The MX-LSTM recursion equation is:

$$\mathbf{h}_t^{(i)} = LSTM\left(\mathbf{h}_{t-1}^{(i)}, \mathbf{e}_t^{(x,i)}, \mathbf{e}_t^{(a,i)}, \mathbf{e}_t^{(H,i)}, \mathbf{W}_{\text{LSTM}}\right). \quad (5)$$

The hidden state of the LSTM model projects onto the four dimensional space, representing the Gaussian multi-variate distribution $\mathcal{N}(\mu_t^{(i)}, \mathbf{\Sigma}_t^{(i)})$, as follows:

$$[\mu_t^{(i)}, \hat{\mathbf{\Sigma}}_t^{(i)}] = \mathbf{W}_o \mathbf{h}_{t-1}^{(i)}, \quad (6)$$

where $\mu_t^{(i)} = [\mu_t^{(x,i)}, \mu_t^{(y,i)}, \mu_t^{(a_x,i)}, \mu_t^{(a_y,i)}]$, $\mathbf{\Sigma}_t^{(i)}$ contains the covariances among the (x, y) coordinate distributions of the tracklets and the vislets, and $\hat{\mathbf{\Sigma}}_t^{(i)}$ is its vectorized version. The

distribution is then sampled to generate the joint prediction of tracklets and vislet points $[\hat{\mathbf{x}}_t, \hat{\mathbf{a}}_t]$, allowing us to simultaneously forecast trajectries and head poses.

At training time, we estimate the weights of the LSTM by minimizing the multivariate Gaussian log-likelihood for the each trajectory. The loss function is

$$L^i(\mathbf{W}_x, \mathbf{W}_a, \mathbf{W}_H, \mathbf{W}_{\text{LSTM}}, \mathbf{W}_o) = -\sum_{T_{obs}+1}^{T_{pred}} log\left(P([\mathbf{x}_t^{(i)}, \mathbf{a}_t^{(i)}], \mu_t^{(i)}, \mathbf{\Sigma}_t^{(i)})\right), \quad (7)$$

where T_{obs} is the last frame of the observation period, while $T_{obs}+1, \ldots, T_{pred}$ are the time frames for which we provide a prediction. The loss of Eq. (7) is minimized over all the training sequences. To prevent overfitting, we additionally include an l_2 regularization term.

4.4 MX-LSTM optimization

As shown in Eq. (7), the optimization procedure provides the weight matrices of the MX-LSTM, which in turn produces the set of Gaussian parameters, including the full covariance $\mathbf{\Sigma}$. The latter is needed to enforce the LSTM in encoding the relations among the (x, y) coordinate distributions of tracklets and vislets, which we already discussed in Sec. 2. In principle, one may have simply captured the correlation between the walking direction and head pose in order to model drifts in the trajectory, but we are interested in letting the MX-LSTM analyze also how the head pose (pan angle) influences the length of the spatial step, that is the velocity. In other words, we want the MX-LSTM to be able to capture whether a particular head pose dynamics could accelerate or slow down the motion, thus letting the machine forecast the joint behavior.

The estimation of a full covariance matrix as the result of an optimization procedure over a generic objective function, like the log-likelihood of (7), is a difficult numerical problem [61]. The main reason is that one must guarantee that the resulting estimate is a proper covariance matrix, *i.e*. a positive semi-definite (p.s.d.) matrix. For this reason, LSTMs with log-likelihood loss functions over Gaussian distributions have been restricted so far to two dimensions, using a simple Gaussian [3], or mixture of Gaussian distributions. The 2×2 covariance matrices have been obtained by optimizing the scalar correlation index $\rho_{x,y}$, which becomes the covariance term of $\mathbf{\Sigma}$ with $\sigma_{x,y} = \rho_{x,y}\sigma_x\sigma_y$ [29].

In case of higher dimensional problems, pairwise correlation terms cannot be optimized for building $\mathbf{\Sigma}$, since the optimization process for each correlation term is independent from each other. At the same time, the positive-definiteness is a simultaneous constraint on multiple variables [62]. In practice, if we consider three variables x, y and z, learning $\rho_{x,y}$ and $\rho_{x,z}$ are two independent procedures, despite that they act on the common distribution over x. This lacks of coordination generates matrices far from being p.s.d. and thus requiring a further correction procedure, It usually consists of projecting the estimated matrix into the closest p.s.d. matrix based on a cost function of the Frobenious norm [12], [39]. This procedure is very expensive [61], and difficult to be embedded into the LSTM optimization process [20], where nonlinearities due to the embedding weights make the analytical derivation hard to formulate. So far, there is not any LSTM loss that involved full covariances of dimension higher than 2.

Our solution involves unconstrained optimization; we use an appropriate Cholesky parameterization of the matrix to be learned

Fig. 3. VFOA pooling: For a given subject, he will try to avoid collision with the people who are inside his view frustum (blue circle). Others (red circle), will not influence his trajectory as they are no in his view frustum.

that enforces the positive semi-definite constraint, dramatically improving the convergence properties of the optimization algorithm [62]. We refer the interested reader to [14] for more details on how to cope with distance measures on covariance matrices. Let us consider $\mathbf{\Sigma}$ a semi-definite positive $n \times n$ (in our case, $n = 4$) covariance matrix. Since $\mathbf{\Sigma}$ is symmetric by definition, only $n(n+1)/2$ parameters are required to represent it. The Choleski factorization is given by:

$$\mathbf{\Sigma} = \mathbf{L}^T\mathbf{L}, \quad (8)$$

where $\mathbf{L}$ is a $n \times n$ upper triangular matrix. The optimization process focuses on finding the $n(n+1)/2$ distinct scalar values for $\mathbf{L}$, which we then solve for the covariance, as for Eq. (8). The main problem with the Cholesky factorization is non-uniqueness: any matrix obtained by multiplying a subset of the rows of $\mathbf{L}$ by -1 is still a valid solution. As a consequence, non-uniqueness makes the problem ill-posed and hinders optimization convergence. The simplest way to enforce the matrix $\mathbf{L}$ to be unique is to add the constraint that all the diagonal elements must be positive. To this end, the Log-Cholesky parameterization [62] assumes that the values found by the optimizer of the main covariance diagonal are the log of the values of $\mathbf{L}$. Formally, the values found by the optimizer can be written as:

$$\theta_L = \begin{bmatrix} \log l_{1,1} & l_{1,2} & l_{1,3} & l_{1,4} \\ 0 & \log l_{2,2} & l_{2,3} & l_{2,4} \\ 0 & 0 & \log l_{3,3} & l_{3,4} \\ 0 & 0 & 0 & \log l_{4,4} \end{bmatrix}. \quad (9)$$

In practice, after the estimation of $\mathbf{W}_x$, $\mathbf{W}_a$, $\mathbf{W}_H$, $\mathbf{W}_{\text{LSTM}}$, $\mathbf{W}_o$ parameters, the values of θ_L are extracted by

$$[\mu_t^{(i)}, \hat{\theta_L}_t^{(i)}] = \mathbf{W}_o\mathbf{h}_{t-1}^{(i)}, \quad (10)$$

where $\hat{\theta_L}$ is the vectorized version of θ_L. Then, the diagonal values of θ_L are exponentiated to form $\mathbf{L}$ and obtaining $\mathbf{\Sigma}$ through Eq. (8).

5 Experiments

To validate the proposed approach we perform both qualitative and quantitative evaluations. We report experiments on two public datasets, namely *UCY* [52] and *TownCentre* [10] datasets. We

TABLE 1
Mean and Final Average Displacement errors (in meters) for all the methods on all the datasets. The first 6 columns are the comparative methods and our proposed model trained and tested with GT annotations. MX-LSTM-HPE is our model tested with the output of a real head pose estimator [37]. The last 3 columns are variations of our approach trained and tested on GT annotations.

Metric	Dataset	SF [84]	LTA [60]	Vanilla LSTM [3]	Social LSTM [3]	Social GAN [31]	MX-LSTM	*MX-LSTM-HPE*	Individual MX-LSTM	NoFrustum MX-LSTM	BD-MX-LSTM
MAD	Zara01	2.88	2.74	0.90	0.68	*0.48*	**0.59**	*0.66*	0.63	0.63	0.60
	Zara02	2.32	2.23	1.09	0.63	*0.44*	**0.35**	*0.37*	0.72	0.36	0.41
	UCY	2.57	2.49	0.67	0.62	*0.65*	**0.49**	*0.55*	0.53	0.51	0.54
	TownCenter	9.35	9.14	4.62	1.96	*1.60*	**1.15**	*1.21*	2.09	1.70	1.40
FAD	Zara01	5.55	5.55	1.85	1.53	*1.04*	**1.31**	*1.43*	1.37	1.40	1.51
	Zara02	4.35	4.35	2.15	1.43	*0.95*	**0.79**	*0.82*	1.56	0.84	1.00
	UCY	4.62	4.66	1.39	1.40	*1.36*	**1.12**	*1.20*	1.16	1.15	1.23
	TownCenter	16.01	16.08	8.26	3.96	*3.50*	**2.30**	*2.38*	4.00	3.40	2.90

compare our model with one baseline, *i.e*. a standard LSTM model that only accounts for pedestrian positions (Vanilla LSTM), and four state-of-the-art approaches: Social Force model (SF) [84], Linear Trajectory Avoidance (LTA) [60], Social LSTM (S-LSTM) [3] and Social GAN [31]. Here we also investigate three variations of the MX-LSTM model to capture the net contributions of the different parts that characterize our approach. Moreover, we investigate the effect of changing the observation period and the forecasting horizon, illustrating how head pose plays a pivotal role for the long term forecasting. Lastly, we analyze whether one can substitute the ground-truth head pose information with more accessible proxies, such as the pace direction or head pose estimates, as provided by a detector. On a qualitative evaluation, we show the interplay between tracklets and vislets that the MX-LSTM has learnt.

5.1 Implementation details

We implemented the MX-LSTM model and all models of the ablation study in Tensorflow. All models have been trained with learning rate of 0.005 along with the RMS-prop optimizer. We set the embedding dimension for spatial coordinates and vislets to 64 and the hidden state dimension is D = 128. We compute the social pooling on a grid of 32×32 cells (3), corresponding to 4 meters. The view frustum aperture angle has been cross-validated on the training partition of the TownCentre and kept fixed for the remaining trials ($\gamma = 40^{\circ}$), while the depth d is simply bounded by the social pooling grid. Training and testing has been accomplished with a GPU NVIDIA GTX-1080 for all evaluations.

5.2 Evaluation Protocol

We report experiments on two public datasets, namely *UCY* [52] and *TownCentre* [10] datasets. The UCY dataset is composed of three sequences (*Zara01*, *Zara02*, and *UCY*), taken in public spaces from top-view. In Table 2, the statistics for each dataset are compared. For all sequences, the manual annotation of the people position and head pose are available (we have annotated the head pose for the TownCenter and made them available at https://github.com/hasanirtiza/MX-LSTM/blob/master/data/).

The evaluation protocol follows the standard procedure for trajectory forecasting that is used in the literature [3], [60]. We first downsample the videos at 2.5fps, then we observe tracklets and vislets for 8 frames, and we predict both locations and head poses for the following 12 time steps. The observation period is $3.2s$ and the forecasting horizon is $4.8s$. Experiments with different time horizons are reported in the ablation study (Sec. 5.4). According to the standard protocol, we use annotations during the observation period. Since we use additional information with respect to most of the related approaches (*i.e*. head poses), we perform an evaluation with the output of a real head pose estimator as well (Sec. 5.4).

TABLE 2
Dataset Statistics

Dataset	Number of frames	Number of pedestrians	Pedestrians per frame	Average trajectories
Zara01	8,670	148	6	339
Zara02	10.513	204	9	467
UCY	5,405	434	32	404
TownCentre	4,500	230	16	310

For the three UCY sequences we train three models, where we use two sequences for training and the remaining for testing. For the TownCentre dataset, the model has been trained and tested on the provided data splits.

Regarding the evaluation metrics of the trajectory forecasting, we consider the *Mean Average Displacement* (MAD) error, *i.e*. the average Euclidean distance between all the predicted and ground-truth pedestrian locations. The *Final Average Displacement* (FAD) error, *i.e*. the Euclidean distance between the last predicted location of each trajectory and the corresponding manually annotated point, is employed as well. Lastly, we evaluate the performance of the head pose predictions in terms of mean angular error e_{α}, which is the mean absolute difference between the estimated pose and the annotated ground truth.

5.3 Comparison with Prior Art

We compare our model against a baseline Vanilla LSTM model, which only uses pedestrian positions, and four state-of-the-art approaches: Social Force model (SF) [84], Linear Trajectory Avoidance (LTA) [60], Social LSTM (S-LSTM) [3] and Social GAN [31].

Note that the Social GAN [31] uses ground-truth trajectories during the prediction interval. At test time, the Social GAN [31] model predicts 20 trajectories and uses the L_2 distance w.r.t. the ground-truth trajectory to select the best one. Although this protocol makes the comparison with all other approach unfair, we include it in the results for the sake of completeness.

Comparative results are reported in Table 1. The MX-LSTM outperforms the state-of-the-art methods across all sequences on both metrics, except for Zara01, where it underperforms the Social GAN. Overall, MX-LSTM achieves an average improvement of 23.3% over the second best performer Social GAN. The highest relative gain is achieved in the UCY sequence and TownCentre dataset, where we achieve a MAD error of 0.49 and 1.15

respectively, improving on Social GAN by 24% and 28% respectively. We explain the larger relative improvement by the increased difficulty of the complex non-linear people paths, in which case the visual attention turns out an important cue. In UCY and TownCenter, people stand in conversational groups, others walk by closely, while some of them slow down to look at the shop windows. We provide quantitative examples of these complex motions in Fig. 1.

Note that some of the evaluated methods require additional input data: both SF and LTA require the destination point of each individual, while SF additionally requires the social group annotations. Ours uses the manually labelled (ground-truth) head poses, which are provided to the algorithm (only) in the observation period (before the forecast). We discuss in the next subsection whether this manual annotation is really needed.

5.3.1 Effect of head pose estimator

Here we analyze the effect on performance, at inference time, of adopting a head pose estimation algorithm [37] during the observation period (prior to forecasting), instead of the ground-truth head poses.

We automatically estimate the head bounding box given the feet positions on the floor plane, assuming an average person being 1.80m tall. Then, we apply the head pose estimator of [37] that provides continuous angles for the pan orientation. At inference time, this data is used as input to this variant, which we name "MX-LSTM-HPE".

Results in Table 1 illustrate that the performance of MX-LSTM-HPE is in average 9% worse than MX-LSTM. The importance of the head pose estimate quality for forecasting is therefore notable, which makes future research on head pose an indispensable requirement. Note from Table 1 that the results of MX-LSTM-HPE are still better than other techniques across all sequences, with the exception of Social GAN [31], outperforming our approach on the UCY sequence.

5.4 Ablation Study

We analyse the net contribution of different parts of the proposed approach by investigating three variations of our model: namely *Block-Diagonal*, *NoFrustum* and *Individual* MX-LSTM.

Block-Diagonal MX-LSTM (BD-MX-LSTM): This studies the importance of estimating full covariances to understand the interplay between tracklets and vislets, rather than modelling each of them as a separate probability distribution. Essentially, instead of learning the 4×4 full covariance matrix $\mathbf{\Sigma}$, BD-MX-LSTM estimates two separate bidimensional covariances $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_a$ for the trajectory and the vislet modeling, thus neglecting the cross-stream covariance. Each 2×2 covariance is estimated employing two variances σ_1, σ_2 and a correlation terms ρ as presented in [29]. The equations that differ from the proposed MX-LSTM are Eq. (7) and Eq. (10), which become:

$$L^i(\mathbf{W}_x, \mathbf{W}_a, \mathbf{W}_H, \mathbf{W}_{\text{LSTM}}, \mathbf{W}_o) = - \sum_{T_{obs}+1}^{T_{pred}} log\left(P([\mathbf{x}_t^{(i)}]^T | \boldsymbol{\mu}_t^{(x,i)}, \boldsymbol{\sigma}_t^{(x,i)}, \rho_t^{(x,i)}\right) + log\left(P([\mathbf{a}_t^{(i)}]^T | \boldsymbol{\mu}_t^{(a,i)}, \boldsymbol{\sigma}_t^{(a,i)}, \rho_t^{(a,i)}\right), \quad (11)$$

where $\boldsymbol{\mu}_t^{(x,i)} = [{}_x\mu_t^{(x,i)}, {}_y\mu_t^{(x,i)}]$ and the same apply for the variance vector;

$$[\boldsymbol{\mu}_t^{(x,i)}, \boldsymbol{\sigma}_t^{(x,i)}, \rho_t^{(x,i)}, \boldsymbol{\mu}_t^{(a,i)}, \boldsymbol{\sigma}_t^{(a,i)}, \rho_t^{(a,i)}]^T = \mathbf{W}_o \mathbf{h}_{t-1}^{(i)}. \quad (12)$$

NoFrustum MX-LSTM:Tthis variant reduces MX-LSTM to the social pooling of [3], i.e. pooling for hidden states $\{\mathbf{h}_t^j\}$ from the entire area around each individual. NoFrustum MX-LSTM neglects the visual frustum of attention and does not select the people to pool from based on it. Also people behind the person would therefore influence the next step forecasting.

Individual MX-LSTM: In this case, no social pooling is taken into account. In more detail, the embedding operation of Eq. (4) is removed, and the weight matrix $\mathbf{W}_H$ vanishes. In practice, this variant learns independent models for each person, each one considering the tracklet and vislet points.

The last three columns of Table 1 report numerical results for the three MX-LSTM variants. The main facts that emerge are: 1) the highest variations are with the Zara02 sequence, where MX-LSTM doubles the performances of the worst approach (Individual MX-LSTM); 2) the worst performing is in general Individual MX-LSTM, showing that social reasoning is indeed needed; 3) social reasoning is systematically improved with the help of the vislet-based view-frustum; 4) full covariance estimation has a role in pushing down the error which is already small with the adoption of vislets.

Summarizing the results so far, having vislets as input allows to definitely increase the trajectory forecasting performance. Vislets should be used to understand social interactions with social pooling, by building a view frustum that tells which are the people currently observed by each individual. All of these features are effectively and efficiently implemented within MX-LSTM. Note in fact that the training time is not affected by whether social pooling is included or not.

Again, although the complete method always outperforms all the competitors, the highest improvement is on the TownCentre sequence. In our opinion this is due to the different level of complexity in the data, indeed most of the trajectories in UCY sequences are relatively linear, with poor social interactions, while in TownCentre there are many interactions, such as forming and splitting groups and crossing trajectories. For the same reason, this is the dataset where the introduction of the view frustum in the pooling of social interactions gives the highest benefits. By contrast, in all other sequences but Zara01, decoupling the covariance matrix into a block diagonal matrix neglecting the interplay of position and gaze (BD-MX-LSTM) leads to a sensitive increase in the prediction error; this proves the tight relation between the head orientation and the motion of an individual.

5.5 Head Pose Forecasting

Our MX-LSTM model also provides a forecast of the head pose of each individual at each frame, for the first time. We evaluate the performances of this estimation in terms of mean angular error e_α, i.e. the mean absolute difference between the estimated pose (angle $\alpha_{t,\cdot}$ in Fig. 2c) and the annotated ground truth. e_α expresses how much the direction in which an individual is looking at a particular time instant is different from the true one. This error measure is independent from the error in the predicted position. In other words, e_α measures the error in the gaze forecasting.

Table 3 reports numerical results of the static head pose estimator [49] (HPE), the proposed model fed with manually annotated head poses (MX-LSTM) and with the output of HPE (MX-LSTM-HPE) during the observation period. In all the cases our forecast output is comparable with the one of HPE, but in our

TABLE 3
Mean angular error (in degrees) for the state-of-the-art head pose estimator [37], and our model fed with manual annotations (MX-LSTM) and estimated values (MX-LSTM-HPE).

Metric	HPE [37]	MX-LSTM	MX-LSTM-HPE
Zara01	14.29	12.98	17.69
Zara02	20.02	20.55	21.92
UCY	19.90	21.36	24.37
TownCentre	25.08	26.48	28.55

TABLE 4
Mean Average Displacement (MAD) error when changing the forecasting horizon. Observation interval is kept constant at 8 frames.

Dataset	Forecasting horizon	Vanilla LSTM	Social LSTM	MX-LSTM	Individual MX-LSTM
Zara 01	**H = 12**	0.90	0.68	**0.59**	0.72
	H = 16	1.21	1.00	**0.87**	1.05
	H = 20	1.70	1.43	**1.21**	1.44
	H = 24	2.30	1.94	**1.55**	1.85
	H = 28	3.07	2.35	**1.92**	2.47
	H = 32	4.11	2.85	**2.40**	3.14
Zara 02	**H = 12**	1.09	0.63	**0.35**	0.63
	H = 16	1.62	0.90	**0.53**	1.09
	H = 20	2.19	1.24	**0.71**	1.43
	H = 24	2.75	1.59	**0.90**	1.83
	H = 28	3.31	2.00	**1.16**	2.25
	H = 32	3.86	2.41	**1.40**	2.67
UCY	**H = 12**	0.67	0.62	**0.49**	0.53
	H = 16	0.90	0.90	**0.70**	0.77
	H = 20	1.19	1.08	**0.95**	1.01
	H = 24	1.52	1.36	**1.22**	1.27
	H = 28	1.87	1.66	**1.50**	1.53
	H = 32	2.24	1.99	**1.80**	1.83

case we do not use appearance cues – i.e. we do not look at the images at all. In the case of Zara01, the MX-LSTM is even better that the static prediction, which highlights the forecasting power of our model. In our opinion, this is due to the fact that in this sequence trajectories are mostly linear and that people are walking fast, with their heads mostly aligned with the direction of motion. When providing the MX-LSTM model with the estimations during the observation period, the angular error increases, as expected, but the error remains limited.

5.6 Time Horizon Effect

To investigate how MX-LSTM performs for longer time horizons we conduct an experimental evaluation where we increment the prediction interval from 12 (standard evaluation protocol) to 32 frames with a step size of 4, keeping the observation interval fixed at 8 frames. We evaluated approaches on UCY, Zara01 and Zara02, since most trajectories on TownCenter last less than 24 frames. We use MAD to report the error. As shown in Table 4, MX-LSTM is well capable of handling longer time horizons. MX-LSTM outperforms all other approaches on all prediction interval, which demonstrates its robustness. Based on these results, we argue that reasoning on the head pose becomes even more important when forecasting in the longer term. Overall, the ranking is preserved and MX-LSTM remains the best performer. Additionally, in Table 5 we also evaluated approaches that are not causal and require ground-truth information during inference time as well. It can be seen that although Social GAN [31] relies on ground-truth information to select the best track during inference, it outperforms MX-LSTM only on Zara01 dataset.

We varied the observation interval, in order to understand how many frames are necessary to learn a meaningful representation of the trajectory, . Table 6 reports numerical results of an experiment

TABLE 5
Comparison of MX-LSTM against techniques which leverage ground-truth information from future frames. Mean Average Displacement (MAD) error when changing the forecasting horizon. Observation interval is kept constant at 8 frames. Note that this comparison is unfair to MX-LSTM, which only uses information from past frames.

Dataset	Forecasting horizon	Social GAN	LTA	SF	MX-LSTM
Zara 01	**H = 12**	**0.48**	2.74	2.88	0.59
	H = 16	**0.68**	3.60	3.65	0.87
	H = 20	**0.94**	4.20	4.21	1.21
	H = 24	**1.26**	4.60	4.61	1.55
	H = 28	**1.66**	4.70	4.74	1.92
	H = 32	**2.20**	4.74	4.82	2.40
Zara 02	**H = 12**	0.44	2.23	2.32	**0.35**
	H = 16	0.60	3.70	3.80	**0.53**
	H = 20	0.76	4.15	4.20	**0.71**
	H = 24	0.95	4.30	4.37	**0.90**
	H = 28	1.17	4.58	4.66	**1.16**
	H = 32	1.43	4.00	4.91	**1.40**
UCY	**H = 12**	0.65	2.49	2.57	**0.49**
	H = 16	0.97	3.17	3.17	**0.70**
	H = 20	1.22	4.20	4.18	**0.95**
	H = 24	1.47	4.48	4.52	**1.22**
	H = 28	1.72	4.60	4.68	**1.50**
	H = 32	1.98	4.88	4.87	**1.80**

TABLE 6
Mean Average Displacement (MAD) error when changing the observation period. Forecasting horizon is kept constant at 12 frames.

Dataset	Observation period	Vanilla LSTM	Social LSTM	MX-LSTM	Individual MX-LSTM
Zara 01	**O = 1**	1.62	**0.89**	0.96	1.43
	O = 4	0.90	0.69	**0.64**	0.79
	O = 8	0.90	0.68	**0.59**	0.72
	O = 12	0.90	0.68	**0.59**	0.68
	O = 16	0.90	0.68	**0.59**	0.60
Zara 02	**O = 1**	1.65	1.13	**0.85**	1.35
	O = 4	1.17	0.74	**0.48**	0.84
	O = 8	1.09	0.63	**0.35**	0.63
	O = 12	1.01	0.63	**0.35**	0.63
	O = 16	0.99	0.63	**0.33**	0.62
UCY	**O = 1**	0.82	0.71	**0.62**	0.88
	O = 4	0.65	0.63	**0.49**	0.59
	O = 8	0.67	0.62	**0.49**	0.53
	O = 12	0.65	0.60	**0.48**	0.52
	O = 16	0.63	0.60	**0.48**	0.52

where we kept the forecasting horizon fixed at 12 frames, and varied the observation period form 1 to 16 frames with the step size of 4 frames. An observation period of 1 frame means we try to predict trajectories based only on a static observation of the individual, with no previous history taken into account. Results prove that one frame is not enough for all the methods under analysis. Despite this, the ranking of different approaches is maintained throughout all the experiments, with the only exception of Zara01 sequence with O=1, where Social LSTM outperforms competitors. Interestingly, a rapid drop in error of about 30% is obtained by observing 4 frames instead of 1. Furthermore, 8 frames are enough for the approaches to learn the overall shape of the trajectory in order to predict for the next 12 frames, as the error drop from observing 8 frames to 16 frames is below 1%.

Finally, in order to understand in more depth how different methods perform for long term forecasting, we kept the observation interval constant at 16 frames and test increasing forecasting horizons. Table 7, further validates the fact that 8 frames are sufficient for the LSTM approach to learn the representation of the trajectory. MX-LSTM is still the best performer but the error drop from observing 8 to observing 16 frames is negligible in long term forecatsing as well. This effect speaks about the

TABLE 7
Mean Average Displacement (MAD) error when changing the forecasting horizon. Observation interval is kept constant at 16 frames.

Dataset	Prediction interval	Vanilla LSTM	Social LSTM	MX-LSTM	Individual MX-LSTM
Zara 01	**Pred = 16**	1.25	1.05	**0.88**	0.90
	Pred = 20	1.27	1.46	**1.19**	1.26
	Pred = 24	1.78	1.88	**1.57**	1.64
	Pred = 28	2.39	2.37	**1.93**	2.01
	Pred = 32	3.09	3.00	**2.32**	2.57
Zara 02	**Pred = 16**	1.31	0.88	**0.49**	0.95
	Pred = 20	1.87	1.24	**0.67**	1.28
	Pred = 24	2.50	1.61	**0.87**	1.65
	Pred = 28	3.19	2.05	**1.11**	2.04
	Pred = 32	3.87	2.53	**1.35**	2.42
UCY	**Pred = 16**	1.02	0.80	**0.71**	0.72
	Pred = 20	1.42	1.06	**0.95**	1.01
	Pred = 24	1.87	1.34	**1.2**	1.40
	Pred = 28	2.37	1.67	**1.46**	1.50
	Pred = 32	2.92	2.21	**1.80**	1.90

TABLE 8
MAD errors on the different datasets

Dataset	MX-LSTM	MX-LSTM-HPE (Train and Test)	Pace-MX-LSTM
Zara01	0.59	0.68	0.69
Zara02	0.35	0.51	0.73
UCY	0.49	0.58	0.59
Town Centre	1.15	1.43	1.50

capability of LSTM-based approaches. The performance already starts to saturate at 8 frames and adding more information does not bring the expected gain. In our view, this highlights the temporal modelling as one of the performance bottlenecks, on the way to progress in the field.

5.7 Substitutes for Head Pose

In this experiment, we analyze the importance of the head pose and question whether one may substitute it with more accessible proxies, such as the direction of the people pace. In more details, we implement a Pace-MX-LSTM, which uses ground truth step directions instead of the head pose. Table 8 illustrates that having the step direction instead of the head pose downgrades the MX-LSTM, since positional data are already contained in the tracklet and the step direction can be extracted from the previous two positions. In fact, Pace-MX-LSTM gives consistently worse results.

In Table 8, we additionally illustrate the importance of having access to manually annotated head poses during training. to study this aspect, we implemented the MX-LSTM-HPE-Train and Test, where the head pose training data is given by a head-pose detector [37]. As expected, MX-LSTM-HPE-Train and Test underperforms MX-LSTM and MX-LSTM-HPE (MX-LSTM-HPE is still trainned on manually labelled head poses, but it adopts a head pose estimator at inference time). This is especially so on Zara02, where conversational groups make the head pose estimation noisy due to the many partial occlusions. Still, MX-LSTM-HPE-Train and Test remains comparable to prior state-of-the-art methods.

5.8 Qualitative Results

Fig. 4 shows qualitative results on the Zara02 dataset, which was found as the most difficult throughout the quantitative experiments. Fig. 4a presents MX-LSTM results: a group scenario is taken into account, with the attention focused on the girl in the bottom-left corner. In the left column, the green ground-truth prediction vislets show that the girl is havign a conversation with the group members, nearly not moving at all, while moving her head around. The magenta curve (Fig. 4a *left*) represents the S-LSTM output, predicting erroneously that the girl would leave the group. This error confirms the problem of competing methods in forecasting the motion of people slowly moving or static, as discussed in Sec. 2. In the central column of Fig. 4a, the observation sequence given to the MX-LSTM is shown in orange (almost static with oscillating vislets). The output prediction (yellow) shows oscillating vislets but no movement, confirming that the MX-LSTM has learnt this particular social behavior. If we provide the MX-LSTM with an artificial observation sequence with the annotated positions (real trajectory) but vislets oriented toward west (third column in Fig. 4a, orange arrows), where no people are present, the MX-LSTM predicts a trajectory slowly departing from the group (cyan trajectory and arrows).

The two rows of Fig. 4b analyze the Individual MX-LSTM, in which no social pooling is taken into account. Here pedestrians are not influenced by the surrounding people, and the forecast motion is only caused by the relationship between the tracklets and the vislets. The first row in Fig. 4b shows three situations in which the vislets of the observation sequence are manually altered to point north (orange arrows), thus orthogonal to the person trajectory. In this case the Individual MX-LSTM predicts a decelerating trajectory drifting toward north (magenta trajectory and vislets), especially visible in the second and third rows. If the observation has the legit vislets (green arrows, barely visible since they are aligned with the trajectory), the resulting trajectory (yellow trajectory and vislets) has a different behavior, closer to the GT (green trajectory and vislets). Similarly, in the second row, we altered vislets to point to South. The prediction with the modified vislets is in black. The only difference is in the bottom left picture: here the observation vislets pointing south are in agreement with the movement, so that the resulting predicted trajectory is not decelerating as in the other cases, but accelerating toward south.

6 Conclusion

We have argued for the importance of people head poses, as encoded in the proposed *vislets*, to forecast their future motion. We have shown that vislets are mostly aligned with the people motion, and therefore useful to forecast it. But when vislets are not aligned with the people motion, then they express the intention of people to change direction. Vislets differ from the current approaches, as most recent LSTM-based forecasting has only considered own and neighboring pedestrian positions. But this is close in spirit to decade-old works using the people desired goals. In this paper, the head pose is however estimated, not provided (e.g. by an oracle).

The use of vislets is enabled by the novel MX-LSTM framework. This jointly "reasons" on tracklets and vislets by means of a multi-variate Gaussian distribution, the covariance of which encodes the interplay of position and head pose. Our proposed log-cholesky parameterization allows its unconstrained optimization by the LSTM backpropagation, and it opens the way to including additional variables (*e.g.* the people belonging to a social group).

Finally, this work has delved into a comprehensive evaluation of the proposed MX-LSTM, including ablation studies on vislets (both estimated and provided as GT), social pooling, view-frustum, observation and prediction time horizons. MX-LSTM provides currently state of the art performance and it is most

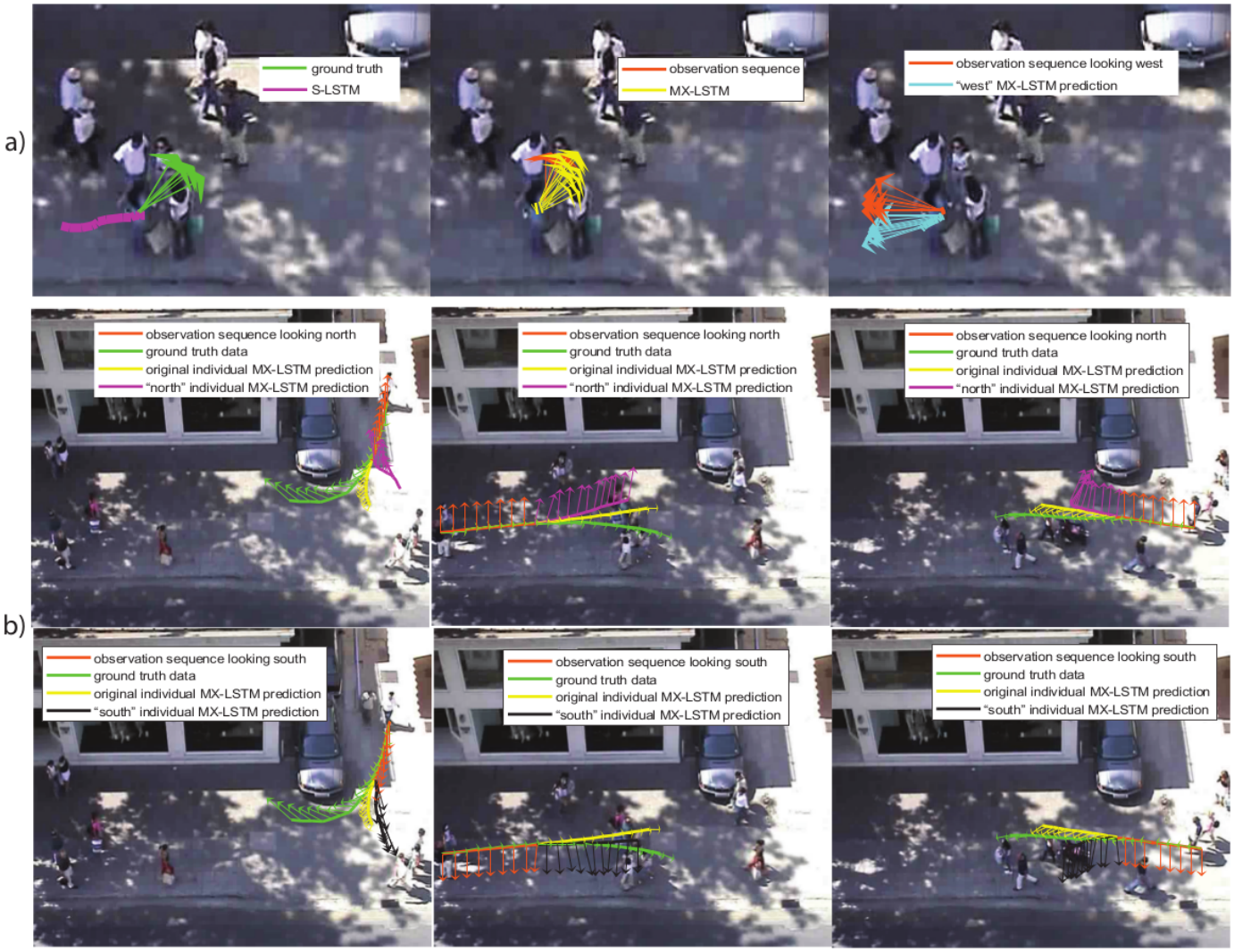

Fig. 4. Qualitative results: a) MX-LSTM b) Ablation qualitative study on Individual MX-LSTM (better in color).

effective when people slow down and look around to change direction, the Achilles heel of other current techniques.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement No. 676455, and has been partially supported by the projects of the Italian Ministry of Education, Universities and Research (MIUR) "Dipartimenti di Eccellenza 2018-2022" and PORFESR 2014-2020 Work Program (Action 1.1.4, project No.10066183).

REFERENCES

[1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.
[2] H. Akaike. Fitting autoregressive models for prediction. *Annals of the institute of Statistical Mathematics*, 21(1):243–247, 1969.
[3] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.
[4] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014.
[5] G. Antonini, M. Bierlaire, and M. Weber. Discrete choice models of pedestrian walking behavior. *Transportation Research Part B: Methodological*, 40(8):667–687, 2006.
[6] G. Antonini, S. V. Martinez, M. Bierlaire, and J. Thiran. Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision*, 69, 2006.
[7] S. O. Ba and J.-M. Odobez. A probabilistic framework for joint head tracking and pose estimation. In *ICPR*, 2004.
[8] S. Becker, R. Hug, W. Hübner, and M. Arens. An evaluation of trajectory prediction approaches and notes on the trajnet benchmark. *arXiv preprint arXiv:1805.07663*, 2018.
[9] B. Benfold and I. Reid. Guiding visual surveillance by tracking human attention. In *BMVC*, 2009.
[10] B. Benfold and I. Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR*, 2011.
[11] A. Bhattacharyya, M. Fritz, and B. Schiele. Long-term on-board prediction of pedestrians in traffic scenes. In *1st Conference on Robot Learning*, 2017.
[12] S. Boyd and L. Xiao. Least-squares covariance matrix adjustment. *SIAM Journal on Matrix Analysis and Applications*, 27(2):532–546, 2005.
[13] J. F. Caminada and W. J. M. van Bommel. Philips engineering report 43, 1980.
[14] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos. Jensen-bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2161–2174, 2012.
[15] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*, pages 215–230. Springer, 2012.
[16] W. Choi and S. Savarese. Understanding collective activities of people

from videos. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1242–1257, 2014.
[17] R. T. Collins, A. J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, 2001.
[18] H. Coskun, F. Achilles, R. Di Pietro, N. Navab, and F. Tombari. Long short-term memory kalman filters: Recurrent neural estimators for pose regularization. In *ICCV*, 2017.
[19] N. Davoudian and P. Raynham. What do pedestrians look at at night? *Lighting Research and Technology*, 44(4):438–448, 2012.
[20] J. E. Dennis Jr and R. B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.
[21] A. D. Dragan, N. D. Ratliff, and S. S. Srinivasa. Manipulation planning with goal sets using constrained trajectory optimization. In *ICRA*, 2011.
[22] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.
[23] J. M. Ferryman, S. J. Maybank, and A. D. Worrall. Visual surveillance for moving vehicles. *International Journal of Computer Vision*, 37(2):187–197, 2000.
[24] S. Fotios, J. Uttley, C. Cheal, and N. Hara. Using eye-tracking to identify pedestrians' critical visual tasks, Part 1. Dual task approach. *Lighting Research & Technology*, 47(2):133–148, 2015.
[25] S. Fotios, J. Uttley, and B. Yang. Using eye-tracking to identify pedestrians' critical visual tasks. part 2. fixation on pedestrians. *Lighting Research & Technology*, 47(2):149–160, 2015.
[26] T. Foulsham, E. Walker, and A. Kingstone. The where, what and when of gaze allocation in the lab and the natural environment. *Vision research*, 51(17):1920–1931, 2011.
[27] N. Gourier, J. Maisonnasse, D. Hall, and J. L. Crowley. Head pose estimation on low resolution images. In *CLEAR*, 2006.
[28] A. Graves. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, 2012.
[29] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.
[30] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.
[31] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. CONF.
[32] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. CONF.
[33] E. T. Hall. *The hidden dimension*. Doubleday & Co, 1966.
[34] Y. D. B. Z. Hang Su, Jun Zhu. Forecast the plausible paths in crowd scenes. In *IJCAI*, 2017.
[35] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, M. Cristani, and F. Galasso. "seeing is believing": Pedestrian trajectory forecasting using visual frustum of attention. In *WACV*. IEEE, 2018.
[36] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani. Mx-lstm: mixing tracklets and vislets to jointly forecast trajectories and head poses. In *CVPR*, 2018.
[37] I. Hasan, T. Tsesmelis, F. Galasso, A. Del Bue, and M. Cristani. Tiny head pose classification by bodily cues. In *ICIP*, 2017.
[38] D. Helbing and P. Molnar. Social force model for. *Physical review E*, 51(5):4282, 1995.
[39] N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear algebra and its applications*, 103:103–118, 1988.
[40] M. Hoai and F. De la Torre. Max-margin early event detectors. *International Journal of Computer Vision*, 107(2):191–202, 2014.
[41] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
[42] S. Huang, X. Li, Z. Zhang, Z. He, F. Wu, W. Liu, J. Tang, and Y. Zhuang. Deep learning driven visual path prediction from a single image. *IEEE Transactions on Image Processing*, 25(12):5892–5904, 2016.
[43] J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognitive psychology*, 43(3):171–216, 2001.
[44] S. R. Jammalamadaka and A. Sengupta. *Topics in circular statistics*, volume 5. World Scientific, 2001.
[45] R. E. Kalman et al. A new approach to linear filtering and prediction problems. *ASME Journal of Basic Engineering*, 1960.
[46] K. Kitani, B. Ziebart, J. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012.
[47] M. Kuderer, H. Kretzschmar, C. Sprunk, and W. Burgard. Feature-based prediction of trajectories for socially compliant navigation. In *Robotics: science and systems*, 2012.
[48] 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.
[49] D. Lee, M.-H. Yang, and S. Oh. Fast and accurate head pose estimation via random projection forests. In *ICCV*, 2015.
[50] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.
[51] N. Lee and K. M. Kitani. Predicting wide receiver trajectories in american football. In *WACV*, 2016.
[52] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. In *Computer Graphics Forum*, 2007.
[53] J. Liu, A. Shahroudy, D. Xu, and G. Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pages 816–833. Springer, 2016.
[54] W. Luo, B. Yang, and R. Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.
[55] W.-C. Ma, D.-A. Huang, N. Lee, and K. M. Kitani. Forecasting interactive dynamics of pedestrians with fictitious play. In *CVPR*, 2017.
[56] J. Mainprice, R. Hayne, and D. Berenson. Goal set inverse optimal control and iterative replanning for predicting human reaching motions in shared workspaces. *IEEE Trans. on Robotics*, 32(4):897–908, 2016.
[57] P. McCullagh and J. A. Nelder. Generalized linear models, no. 37 in monograph on statistics and applied probability, 1989.
[58] B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE Trans. on Circuits and Systems for Video Technology*, 18(8):1114–1127, 2008.
[59] A. E. Patla and J. N. Vickers. How far ahead do we look when required to step on specific locations in the travel path during locomotion? *Experimental brain research*, 148(1):133–138, 2003.
[60] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009.
[61] J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296, 1996.
[62] M. Pourahmadi. Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, pages 369–387, 2011.
[63] M. B. Priestley. *Spectral analysis and time series*. Academic press, 1981.
[64] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(12):1939–1959, 2005.
[65] C. E. Rasmussen. Gaussian processes for machine learning. In *Adaptive Computation and Machine Learning*, 2006.
[66] N. M. Robertson and I. D. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, 2006.
[67] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *ECCV*, 2016.
[68] M. S. 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.
[69] A. Sadeghian, A. Alahi, and S. Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. *arXiv preprint arXiv:1701.01909*, 2017.
[70] N. Srivastava, E. Mansimov, and R. Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.
[71] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel. From gaze to focus of attention. In *VISUAL*, 1999.
[72] H. Su, Y. Dong, J. Zhu, H. Ling, and B. Zhang. Crowd scene understanding with coherent recurrent neural networks. In *IJCAI*, 2016.
[73] L. Sun, Z. Yan, S. M. Mellado, M. Hanheide, and T. Duckett. 3DOF pedestrian trajectory prediction learned from long-term autonomous mobile robot deployment data. *arXiv preprint arXiv:1710.00126*, 2017.
[74] D. Tosato, M. Spera, M. Cristani, and V. Murino. Characterizing humans on riemannian manifolds. *IEEE TPAMI*, 35(8):1972–1984, 2013.
[75] P. Trautman and A. Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *IROS*, 2010.
[76] A. Treuille, S. Cooper, and Z. Popović. Continuum crowds. *ACM Transactions on Graphics (TOG)*, 25(3):1160–1168, 2006.
[77] P. Vansteenkiste, G. Cardon, E. D'Hondt, R. Philippaerts, and M. Lenoir. The visual control of bicycle steering: The effects of speed and path width. *Accident Analysis & Prevention*, 51:222–227, 2013.
[78] D. Varshneya and G. Srinivasaraghavan. Human trajectory prediction using spatially aware deep attention models. In *NIPS*, 2017.
[79] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.
[80] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical

models for human motion. *IEEE TPAMI*, 30(2):283–298, 2008.
[81] C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models*, pages 599–621. Springer, 1998.
[82] Z. Xing, J. Pei, G. Dong, and P. S. Yu. Mining sequence classifiers for early prediction. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 644–655. SIAM, 2008.
[83] H. Xue, D. Q. Huynh, and M. Reynolds. Ss-lstm: A hierarchical lstm model for pedestrian trajectory prediction. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1186–1194. IEEE, 2018.
[84] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *CVPR*, 2011.
[85] S. Yi, H. Li, and X. Wang. Understanding pedestrian behaviors from stationary crowd groups. In *CVPR*, 2015.
[86] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.
[87] 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.

Irtiza Hasan is a research associate at Inception Institute of Artificial Intelligence (IIAI). His interests involve object detection and tracking, activity forecasting and pose estimation. He obtained his PhD from University of Verona, Italy in 2019. During his PhD, he was also a scientist at Computer Vision Department at OSRAM (Munich, Germany). His PhD research focused on smart lighting and applying computer vision techniques to understand and forecast people activities.

Francesco Setti is an Assistant Professor at the Department of Computer Science of the University of Verona working on the EU-H2020 project SARAS, and Associate Researcher of the Institute of Cognitive Science and Technology (ISTC-CNR). His research interests are in the application of machine learning and artificial intelligence techniques for industrial applications, with particular attention to the emerging fields of collaborative robotics and reinforcement learning for situation awareness decision making.

Theodore Tsesmelis is currently a computer vision researcher at the Computer Vision Department at OSRAM GmbH in Germany. Currently he is applying research and development regarding smart workspace management with the use of computer vision and machine learning. Other research interests involve reconstruction and recognition of 3D scenes, the scene material and object properties, and the estimation of lighting propagation.

Vasileios Belagiannis is Junior-Professor at University of Ulm. He holds a degree in engineering from Democritus University of Thrace and M.Sc. in Computational Science and Engineering from TU München. He completed his doctoral studies at TU München and continued as post-doctoral research assistant at University of Oxford (Visual Geometry Group). Prior to joining University of Ulm, he was conducting research at OSRAM in Germany. His research is focused on deep learning, machine learning and computer vision, including applications from autonomous driving.

Sikandar Amin is a senior computer vision scientist at OSRAM in Germany, where he manages multiple AI projects regarding infrastructure sensing platforms for autonomuos driving, and indoor smart lighting applications. His R&D interests include object detection, tracking, re-identification and full body pose estimation. He obtained his PhD from TU München in computer vision. His PhD research includes 2D and 3D human pose estimation for higher level tasks including activity recognition and studying human emotions during dyadic interactions in complex real-world settings.

Alessio Del Bue is a tenured senior researcher at the PAVIS (Pattern Analyisis and computer VISion) Department of the Italian Institute of Technology (IIT) where he is leading the Visual Geometry and Modelling (VGM) Lab. His current research interests are related to 3D scene understanding from images and sound, 3D digitisation technology for Cultural Heritage studies, non-rigid image registration and reconstruction, and sensors/targets localisation and room reconstruction from sound.

Marco Cristani is an Associate Professor (Professore Associato) at Computer Science Department, University of Verona. His main research interests are in statistical pattern recognition (mainly deep learning and generative modeling) and computer vision, with emphasis on social signal processing, *i.e.*, how to model human activities with computer vision tools, following social psychology principia. In particular, he is interested in fashion modeling (clothing parsing, recommendation, attribute learning) and how fashion is related to personality. Other interests are on video surveillance applications, such as human activity modeling, people re-identification and pedestrian detection.

Fabio Galasso heads the Computer Vision Department at OSRAM (Munich, Germany), an international team conducting R&D in artificial intelligence, computer vision and machine learning, in relation to smart lighting applications. Prior to OSRAM, he has conducted research at the University of Cambridge (UK) and at the Max Planck Institute for Informatics (Germany). He received his Master's Degree cum laude from the RomaTre University (Italy) and his PhD from the University of Cambridge (UK).