arXiv:1803.03347v1 [cs.CV] 9 Mar 2018

Tracking by Prediction: A Deep Generative Model for Mutli-Person localisation and Tracking

Tharindu Fernando Simon Denman Sridha Sridharan Clinton Fookes

Image and Video Research Laboratory, Queensland University of Technology (QUT), Australia
{t.warnakulasuriya, s.denman, s.sridharan, c.fookes}@qut.edu.au

Abstract

Current multi-person localisation and tracking systems have an over reliance on the use of appearance models for target re-identification and almost no approaches employ a complete deep learning solution for both objectives. We present a novel, complete deep learning framework for multi-person localisation and tracking. In this context we first introduce a light weight sequential Generative Adversarial Network architecture for person localisation, which overcomes issues related to occlusions and noisy detections, typically found in a multi person environment. In the proposed tracking framework we build upon recent advances in pedestrian trajectory prediction approaches and propose a novel data association scheme based on predicted trajectories. This removes the need for computationally expensive person re-identification systems based on appearance features and generates human like trajectories with minimal fragmentation. The proposed method is evaluated on multiple public benchmarks including both static and dynamic cameras and is capable of generating outstanding performance, especially among other recently proposed deep neural network based approaches.

1. Introduction

Multi-person localisation and tracking is one of the most active research areas in computer vision as it enables a variety of applications including sports analysis [18, 39, 65], robot navigation [10, 11] and autonomous driving [12, 15, 48].

Despite the impact of deep learning across a multitude of computer vision domains in recent years, within the tracking space it has been applied in a somewhat piecemeal manner, with it often used for only a specific part of the tracking pipeline. For example, techniques such as [34, 58, 62] use DCNNs to model subject appearance within a probabilistic tracker. We note that to date, complete deep learning solutions for both localisation and tracking have been limited [41]. We believe this is due to the scarcity of training data which is large enough to train a complete deep neural network based platform, as well as the complex, variable length nature of multi person trajectories. We combine this with a deep tracking framework that utilises Long Short Term Memory networks (LSTMs) to capture pedestrian dynamics in the scene and track objects via predicting it's future trajectory.

In this paper we contribute a novel light weight person detection framework based on Generative Adversarial Networks (GANs) [23], which can be easily trained on the limited data available for multi person localisation. We extend the general GAN framework to temporal sequences and render a probability map for pedestrians in the given sequence. The temporal structure of the proposed GAN allows us to identify pedestrians more effectively, regardless of the motion of other foreground objects in the scene.

As illustrated in [42], multi person tracking consists of two subproblems: data association (i.e assigning a unique identifier to the corresponding targets) and inferring the trajectories of the targets. In most data association paradigms the researchers utilise an appearance based model [3, 9, 53, 60, 61] to re-identify the targets in the next frame. Yet in crowded environments with a high likelihood of occlusions, noisy detections, and poor image resolution, appearance models often fail to generate correct identification of the targets. This results in an un-realistic trajectory generation from the tracking process. To counter this problem we propose a novel tracking framework where the object detections in the next frame are associated with targets via considering their predicted short term and long term trajectories. The trajectory prediction method accounts for the motion of the pedestrian as well as the motion of other people in the local neighbourhood which allows the modelling of the interactions among them. This enables intelligence in the data association process generating human like trajectories even in the presence of occlusions and other image artefacts. To achieve this, we build upon recent advances [16, 41] in pedestrian trajectory modelling approaches and

propose a method to capture relationships with neighbourhood dynamics as well as the long term dependencies within the scene context.

The major contributions of the proposed work can be summarised as follows:

- We introduce a novel pedestrian detection platform based on Generative Adversarial Networks (GANs).
- We develop a robust light weight algorithm for data association in multi person tracking problems with the aid of trajectory prediction.
- We generate human like trajectory estimates via the association of neighbourhood and scene context in the trajectory prediction framework.
- We comprehensively evaluate the proposed models on publicly available benchmarks including videos from both static and dynamic cameras.
- We achieve outstanding performance in the MOT challenge benchmark datasets, especially among deep neural network based approaches.

2. Related work

2.1. Pedestrian detection and localisation

In classical literature, hand engineered feature vectors composed of different pedestrian characteristics are used to train classifiers on image patches. In [31] the authors use a Random Forest classifier trained by boosting where as in [24] the authors integrate different sources of information (i.e. foreground information, object shape) with probabilistic graphical models. This was further extended in [55] where the authors incorporate other parameters for detection such as object position, ground plane parameters and confidence of the detection through a Bayesian network.

With the dawn of deep learning hand engineered feature learning has been superseded by automatic feature learning approaches as they can learn more informative, multiple feature hierarchies. In the first attempts to utilise deep methods for pedestrian detection, authors in [44] built upon the classical model of [13] and used a stack of Restricted Boltzmann Machines. This work was further extended by Ouyang et al. [45] where the authors account for person-to-person relations. In a similar line of work, Sermanet et al. [56] used a combination of features from the last two layers of a CNN for the detection task. The detection is performed in a sliding window fashion where different scales are used for detecting in different granularities.

Most recently, in [25] the authors discuss the importance of the R-CNN pipeline [22] for pedestrian detection. This method scans through image patches at the superpixel level to determine regions of interest for pedestrians before extracting out CNN features. In the next step extracted features are passed to an SVM which classifies the class of each object. Finally it passes through a localisation layer to determine the specific object location. Even though this R-CNN approach renders accurate detections, the networks structure is inherently complex and computationally expensive. It is prohibitive to scan an image at the superpixel level and perform convolution operations within those extracted patches [26]. Single Shot MultiBox Detector (SSD) [37] and You Only Look Once (YOLO) [50] models built upon the RCNN model and propose a simpler network architecture with fewer parameters. Yet when processing videos, the usual approach is to process them framewise, completely discarding the temporal relationships that exist between frames, which can be taken into consideration to help overcome clutter and occlusions.

We build upon the recent success of GANs [17, 23] and propose a method for pedestrian localisation in video frames. We expand the GAN framework to videos, mapping the temporal relationships between frames. This accounts for occlusions and false detections which can appear due to the motion of non-pedestrian objects in the scene.

2.2. Pedestrian tracking

Related literature often handles the data association via recursive Bayesian filters such as Kalman filters [6], particle filters [43] or relying on the appearance based pedestrian characteristics [29, 59, 64] such as position, height etc. Another line of work has emerged where the multi object trajectories are constructed through optimisation strategies [2, 7, 31, 42]. Yet, instead of learning the features for data association in a data driven way, these approaches use hand engineered features, totally relying on the domain knowledge of the composer.

Several attempts have been made to introduce deep learning for data association in tracking. Siamese networks [34, 62] and quadruplet networks [58] have been introduced to perform data association by considering appearance features, yet haven't been able to generate substantial impact when comparing their performance against probabilistic methods such as [31, 42] in public benchmarks.

With recent advances in sequence modelling with deep learning approaches, a data driven method for crowd motion modelling has been proposed in [1], which was further expanded to capture the entire history from the neighbourhood in [16]. In a different line of work authors in [14] have looked into modelling long term dependencies such as trajectory patterns between sequences for the task of trajectory prediction, rather than using dependencies within the sequence. However, these methods were engineered for long term prediction of trajectories as opposed to tracking.

Numerous attempts have been made to transfer these

deep pedestrian motion modelling approaches for data association to tracking. In [53] the authors utilise LSTMs as motion and interaction models, in addition to using an appearance model to cope with the occlusions, noisy detection and appearance changes. They utilise two separate LSTMs as their motion and interaction models. Still their approach needs tedious processing such as occupancy map generation to obtain interaction features. The authors of [41] proposed an LSTM based end-to-end approach for multi target tracking, including initiation and termination of objects, bounding box prediction and data association. Yet they do not consider interactions among objects and the approach results in erroneous and non realistic trajectory generation.

In the proposed model we show how automatic learning of such interactions is possible in a data driven manner. We introduce a coherent architecture for capturing temporal dependencies from multiple information cues derived from the deep trajectory planner in [16]. In addition to the historical behaviour of the pedestrian of interest we capture information from its local neighbourhood as well as contextual information such as intention and group behavioural patterns, which are stored as latent variables in the trajectory planner. We elaborate on how these information streams can be utilised instead of an appearance model when performing data association.

3. Tracking framework

Fig. 1 depicts the proposed tracking framework. We pass each input frame through the GAN generator. This yields a probability map which classifies the likelihood of each pixel of the input frame being a pedestrian or not (see Section 3.1). Then we apply watershed segmentation [57] to segment out each connected component. In order to maintain the trajectory information of each object we retain an object pool in the memory. The segmentation of results of the first frame is used to initialise the object pool. In the proposed framework data association is performed through a trajectory prediction process. The predicted short term trajectory is used for data association where as the long term trajectory prediction is used for trajectory update of the objects to render human like trajectories in the presence of occlusions and other image artefacts. As the final step we update our predictions as well as the object pool by adding newly created objects to the pool and terminating tracking for objects that have not been updated recently. A detailed explanation of the process is given in the following subsections.

3.1. Person localisation through GANs

Generative adversarial networks learn a mapping from a random noise vector z to an output image $y, G : z \rightarrow y$ [23]. The generator (G) is trained to generate data that is indistinguishable from real data, while the discriminator (D) is trained to identify the generated data. This objective can

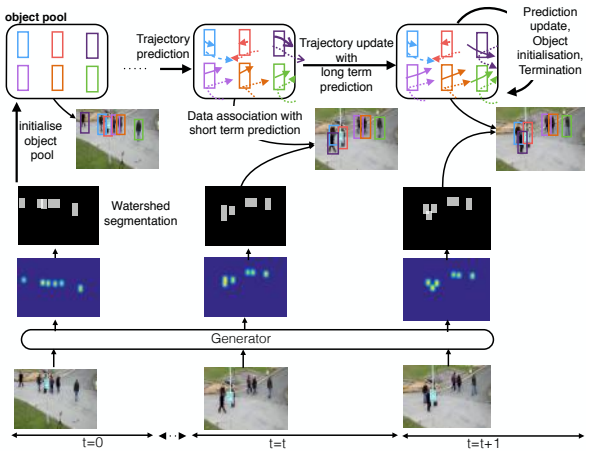


Figure 1. The proposed deep tracking framework: In the first frame of the sequence the object pool is initialised via the person detections generated by passing the input frame through the person detector framework (Section 3.1). We apply watershed segmentation to segment the probability map generated by the generator. Then the trajectory prediction process predicts the short term and long term trajectory predictions where the former is used for data association and the latter is used to update the trajectory of the objects and render human like trajectories in the presence of occlusions and other image artefacts. Finally we update our predictions as well as the object pool by adding newly created objects to the pool and terminating the objects that have not been updated recently.

be written as,

$$L_{GAN}(G,D) = E_{x,y\sim p_{data}(x,y)}[log(D(x,y))] + E_{x\sim p_{data(x)},z\sim p_z(z)}[log(1-D(x,G(x,z)))]. \quad (1)$$

We utilise a generative model to create a probability map for person detection, which classifies the likelihood of each pixel of the input frame being a pedestrian. We extend the general image to image synthesis framework of GAN to video segments, by incorporating an LSTM module between the encoder and decoder of the generator. This allows us to focus on both spatial and temporal dynamics of the different regions in the scene and generate scene specific detection maps for the pedestrians.

Then the loss function of the proposed model for t frames can be written as,

$$L_{GAN}(G,D) = \frac{1}{t}\sum_{t=1}^{t} E_{x_t,y_t\sim p_{data}(x_t,y_t)}[log(D(x_t,y_t))] + \frac{1}{t}\sum_{t=1}^{t} E_{x_t\sim p_{data(x_t)},z_t\sim p_z(z_t)}[log(1-D(x_t,G(x_t,z_t)))]. \quad (2)$$

Let Ck denote a Convolution-BatchNorm-ReLU layer group with k filters. CDk denotes a Convolution-BatchNorm-Dropout-ReLU layer with a dropout rate of

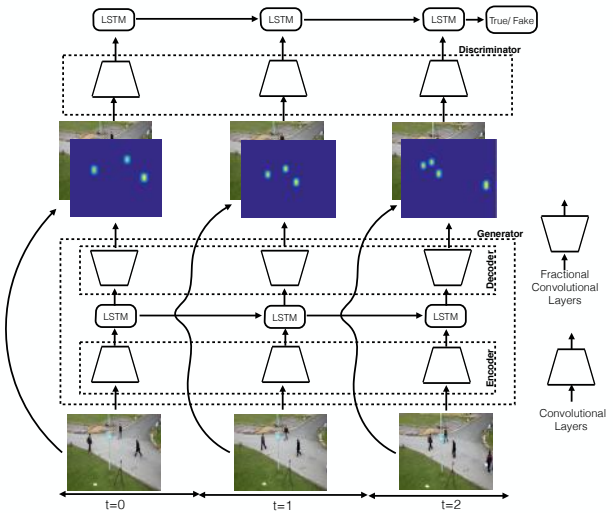


Figure 2. The proposed pedestrian detection framework : Video frames are passed through the encoder of the generative model frame wise. The temporal relationships between the encoder embeddings are mapped through an LSTM layer. Finally the decoder maps these embeddings to a probability distribution which classifies the likelihood of each pixel of the input frame being a pedestrian. The discriminator of the GAN framework takes both the input frame sequence and the generated probability distribution sequence for those frames into consideration and generates a single classification via passing it through an LSTM network.

50%. Then the generator architecture can be written as, Encoder: C64-C128-C256-C512-C512-C512-C512-C512 followed by an LSTM module with 64 hidden units. For the decoder we use CD512-CD1024-CD1024-C1024-C1024-C512-C256-C128. The discriminator architecture is: C64-C128-C256-C512-C512-C512 and finally an LSTM with 64 hidden units. The only difference between the LSTM architecture in the encoder and the decoder is that the former is a sequence-to-sequence LSTM which generates an output at each timestep, where as the latter is a sequence-to-1 LSTM which generates a single output considering the entire sequence [20]. All convolutions are 4 x 4 spatial filters applied with stride 2. Convolutions in the encoder and in the discriminator down sample by a factor of 2, whereas in the decoder they up sample by a factor of 2 (i.e fractional Convolutional layers).

To sample video segments from the input video we use a sliding window and sample segments with a length of 10 frames. The motivation behind choosing this window length is explained in Sec. 3.2. We trained the proposed model with the Adam [30] optimiser, with a batch size of 32 and an initial learning rate of $1 \times e^{-5}$, for 10 epochs.

3.2. Tracking using Trajectory Prediction

An object pool: We retain an object pool in the memory which is initialised from the segmentation results of the first frame. The object pool is associated with a trajectory prediction module which regresses short term and long term trajectories for each object in the pool based on the historic trajectory of that object and neighbouring trajectories. Therefore, each object that resides in the object pool is associated with a probable short term and long term trajectory. For the short term trajectory prediction, we utilise a trajectory and local neighbourhood from the previous 2 frames to predict it's trajectory for the next 2 frames. In contrast, for long term prediction we consider the same attributes for the last 10 frames and predict for the next 10 frames.

The motivation for using both long and short term predictors can be illustrated as follows. In [53] the authors evaluated the occlusion length (in frames) of the MOT validation set and their results show that most occlusions last around 3 frames. Yet there exist around 150 occlusions that last 3-10 frames. These occur largely due to interactions among groups and can be overcome via long term trajectory prediction. Even though we are not using the same dataset, the findings from [53] validates the argument for both short and long term trajectory planning. Furthermore, as pointed out in [42, 53] raw detection outputs can be noisy due to occlusions, false alarms, inaccurate bounding boxes, and missing detections. A data association process which relies merely on the short term trajectory of an object may be overly sensitive to this noise and be unable to account for important pedestrian trajectory dynamics such as collision avoidance, persistence and group motion.

For the trajectory prediction algorithm we build upon the work of [41] to incorporate the motion of the pedestrian of interest and it's neighbourhood. Authors in [16] show that their algorithm can account for both long and short term trajectory planning including interactions among pedestrians and group motion. Therefore we improve the deep tracking framework in [41] with the attributes derived in [16] for trajectory prediction and show how those attributes can be utilised in a deep tracking platform, eliminating the need for separate motion and interaction models as in [53].

Trajectory prediction: We use the trajectory prediction framework of [16] to predict motion. Let the historical trajectory of pedestrian k, from frame 1 to frame T_{obs} be given by,

$$\mathbf{p^k} = [p_1, \ldots, p_{T_{obs}}], \tag{3}$$

where the trajectory is composed of points in a Cartesian grid. Then we pass these historical trajectories through the LSTM encoder of each respective pedestrian to generate its historical embeddings as follows,

$$h_t^k = \text{LSTM}(\mathbf{p}_t^k, h_{t-1}), \tag{4}$$

generating a sequence of historical embeddings,

$$h^k = [h_1^k, \ldots, h_{T_{obs}}^k]. \tag{5}$$

We utilise a soft attention context vector $C_t^{s,k}$ to embed the trajectory information from the pedestrian of interest (k), which can be computed as a weighted sum of hidden states,

$$C_t^{s,k} = \sum_{j=1}^{T_{obs}} \alpha_{tj} h_j^k, \tag{6}$$

and the weight α_{tj} can be computed by,

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{l=1}^{T} exp(e_{tl})}, \tag{7}$$

$$e_{tj} = a(h_{t-1}^k, h_j^k), \tag{8}$$

where the function a is a feed forward neural network for joint training with other components of the system.

The hardwired attention context vector $C^{h,k}$ is used for embedding the neighbouring trajectories. The hardwired weights, denoted by w, can be computed as,

$$w_j^n = \frac{1}{\text{dist}(n,j)}, \tag{9}$$

where $\text{dist}(n,j)$ is the distance between the n^{th} neighbour and the pedestrian of interest at the j^{th} time instant, and w_j^n is the generated hardwired attention weight. When there are N neighbouring trajectories in the local neighbourhood, and h_j^n is the encoded hidden state of the n^{th} neighbour at the j^{th} time instant, then the context vector for the hardwired attention model is defined as,

$$C^{h,k} = \sum_{n=1}^{N} \sum_{j=1}^{T_{obs}} w_j^n h_j^n. \tag{10}$$

The merged context vector, $C_t^{*,k}$, computed by,

$$C_t^{*,k} = \tanh([C_t^{s,k}; C^{h,k}]), \tag{11}$$

is then used to predict the future trajectory for the pedestrian of interest,

$$\mathbf{q}_t = \text{LSTM}(h_{t-1}^k, \mathbf{q}_{t-1}, C_t^{*,k}). \tag{12}$$

Note that $\mathbf{q}_t$ is composed of points in a Cartesian gird. We let $T_{obs} = 3$ for short term predictions and $T_{obs} = 10$ for long term trajectory predictions. After predicting the short term and long term trajectories the predictions are stored in respective objects along with the sequence of context vectors $C^{*,k}$ that is used for long term trajectory prediction. In contrast to [16], where the authors cluster the trajectories and learn separate motion models, we learn one single motion model to predict all trajectories individually. Similar to [16], the hidden state dimensions of all the LSTM embeddings (i.e h^k) are set to be 300 hidden units. Both short and long term trajectory prediction models are pretrained on the dataset of [40], with the Adam optimiser and an initial learning rate of $1 \times e^{-4}$ and batch size of 32 for 100 epochs. We fine-tuned on the training set of the respective dataset with an $1 \times e^{-5}$ learning rate, due to limited data availability preventing us training models from scratch.

Data association with short term trajectory predictions: In the next step we associate each detection (see Section 3.1) with an object from the object pool, if the centroid of the segment lies within a distance threshold to the predicted short term trajectory. The trajectory history of the associated object is updated accordingly.

New object initialisation: If no object in the pool lies within the given threshold, a new object is created and added to the pool.

Trajectory update with long term trajectory prediction: We utilise a long term prediction module to render smooth human like tracking outputs, overcoming the noisy discontinuous nature of the detections. All the objects that reside in the object pool are compared pairwise and merged if they exhibit a similar long term trajectory and if there is similarity between the context vector sequences of the two pedestrians.

Let the two pedestrians be denoted as k_1 and k_2 and the long term trajectory predictions for the period T_{obs+1} to T_{pred} (i.e from 11 to 20) be $p^{k_1} = [p^{k_1}_{T_{obs+1}}, \ldots, p^{k_1}_{T_{pred}}]$ and $p^{k_2} = [p^{k_2}_{T_{obs+1}}, \ldots, p^{k_2}_{T_{pred}}]$ respectively, where p_t^k are points in a 2D Cartesian grid. We denote the context vector sequences from Eq. 11 for the respective pedestrians, for generating the long term trajectory as $C^{*,k_1} = [C_1^{*,k_1}, \ldots, C_{T_{obs}}^{*,k_1}]$ and $C^{*,k_2} = [C_1^{*,k_2}, \ldots, C_{T_{obs}}^{*,k_2}]$. Then the spatial dissimilarity (SD) between the two objects can measured using the Hausdorff distance [67] between the long term trajectory predictions of the objects. We use Hausdorff distance as it is widely used to measure the spatial similarity between the pedestrian trajectories in a surveillance setting [27, 38]. This can be denoted as,

$$SD = \max(d(p^{k_1}, p^{k_2}), d(p^{k_2}, p^{k_1})), \tag{13}$$

where $d(p^{k_1}, p^{k_2}) = \max_{a \in p^{k_1}} \min_{b \in p^{k_2}} ||a - b||$.

In [54] the authors have shown that the cosine-similarity distance measure was more effective for discriminating the hidden states of deep neural networks than traditional methods of using Jacquard similarity or SVMs. Therefore we measure the context dissimilarity (CD),

$$\begin{aligned} CD &= 1 - \frac{C^{*,k_1} \cdot C^{*,k_2}}{||C^{*,k_1}||_2 ||C^{*,k_2}||_2} \\ &= 1 - \frac{\sum_{t=1}^{T_{obs}} C_t^{*,k_1} C_t^{*,k_2}}{\sqrt{\sum_{t=1}^{T_{obs}} (C_t^{*,k_1})^2} \sqrt{\sum_{t=1}^{T_{obs}} (C_t^{*,k_2})^2}} \end{aligned} \tag{14}$$

Table 1. Person detection evaluations on PETS S1L2 dataset

Method	MODA	MODP	Precision	Recall
Peng et al. [47]	0.79	0.74	0.92	**0.91**
Ge et al. [21]	0.75	0.68	0.85	0.89
Chavdarova et al. [8]	0.88	0.75	0.97	**0.91**
Proposed	**0.91**	**0.79**	**0.98**	**0.91**

Table 2. Person detection evaluations on PETS S2L1 dataset

Method	MODA	Precision	Recall
DOR [5]	0.60	0.93	0.87
POM-CNN [5]	0.43	0.90	0.86
Faster R-CNN [52]	0.27	0.50	0.63
Proposed	**0.69**	**0.95**	**0.91**

Then we retain only the older object, and discard the young object from the object pool if the dissimilarities (i.e SD and CD separately) between the two objects are less than specified thresholds, which are evaluated experimentally and available in the supplementary material.

Termination: All the objects that do not get updated for 10 consecutive frames are removed from the pool.

Prediction update: Finally we update the predictions for the next time segment. As depicted in Fig. 1, we repeat this process for all the frames in the given video.

4. Experimental Results

4.1. Person Detector evaluation

We evaluate the proposed person detector on the PETS2009-S1-L2 [19] dataset which is a widely recognised benchmark dataset for pedestrian detection. It contains seven outdoor sequences from seven cameras, with 795 frames for every sequence. Similar to [8, 21, 47] we retained camera view 1 for testing and trained our detector on the other camera views. In addition to the empirical precision and recall estimates of the detector, we also report the Multiple Object Detection Accuracy (MODA) and Multiple Object Detection Precision (MODP) metrics from [28]. MODA accounts for the normalised missed detections and MODP assesses the localisation quality of the true positives.

From the results shown in Tab. 1 our method performs well compared to existing state-of-the-art baselines, which is largely due to the augmented temporal structure of the GAN. This provides the detector the capability to leverage appearance information, and filter out foreground objects by considering both appearance and motion features, performing a precise detection of the humans.

Furthermore, we evaluate our person detector performance on the PETS S2L1 dataset which is one of the videos available for training in 3D MOT 2015 benchmark [35] and compare it against three recent baseline methods. It should be noted that POM-CNN [5] utilises a CNN-based foreground segmentation process while DOR [5] has coupled the CNN with conditional random fields specifically to handle occlusions and imitate the generative-discriminative training approach of GANs. Still the proposed approach outperforms the state-of-the-art methods in all the considered metrics. We were unable to compare the MODP metric as that information for the baselines is not available. It should be noted that the proposed generative model for pedestrian detection has only 32,869,313 parameters for training where as the region proposal network of the Tiny-Yolo [51] has 45,079,472 parameters.

4.2. Tracker evaluations

This section evaluates the proposed deep tracker on the 3D MOT 2015 benchmark [35], which is composed of the PETS09-S2L2 and AVG-TownCentre sequences. For a fair comparison with other baselines we use the provided pedestrian detections. Furthermore, we evaluated the proposed model on ETH Mobile Scene (ETHMS) dataset [11], which is challenging with a busy pedestrian street filmed with a moving stereo camera. It should be noted that we do not use the available camera calibration or depth maps, but rather track the people in the image space.

The reported metrics are the ones suggested in the 3D MOT 2015 benchmark: Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Mostly Tracked targets (MT), Mostly Lost targets (ML), False Positives (FP), False Negatives (FN), ID Switches (IDS), and the number of frames processed in one second (Hz) denoting the speed of a tracking method.

For the results presented in Tab. 3 and Tab. 4 for the 2 sequences of the 3D MOT 2015 benchmark, it can be observed that the proposed system exhibits better performance among all the models for most metrics. Furthermore, the proposed algorithm exhibits a noticeable increase in both MOTA and MT, and reductions in ML and trajectory fragmentations (Frag), compared to other deep neural network approaches. Specifically, we compare the proposed method against state-of-the-art deep neural network based methods. Wang et al. [62] performs data association using both motion and appearance feature cues. The appearance features are learnt using CNNs where as the motion based track similarity is evaluated based on the velocity, which is handcrafted. Yet their method generates lower MOTA and MT values and very high ML values, in both sequences compared to our method. Furthermore, we would like to point that the CNN based appearance feature extraction method has led their architecture to have a high time complexity.

A simpler data association scheme is proposed in [34] where the authors only use appearance features and utilise linear programming to associate the tracklets, which leads to a comparatively higher speed. Even though they are achieving fewer FPs, their tracker misses the majority of the targets resulting a higher FN value and ML percentage, and a lower MT rate. The MOTA value associated with their

Table 3. 3D MOT 2015 results for PETS09-S2L2. Arrows indicate favourable direction of each metric. Best values are printed in bold

Type	Tracker	MOTA ↑	MOTP ↑	MT (%) ↑	ML (%) ↓	FP ↓	FN ↓	IDS ↓	Frag ↓	Hz ↑
Probabilistic	Klinger et al. [31]	**57.6**	65.5	**28.6**	4.8	805	3049	231	245	0.1
	Klinger et al. [32]	55.5	63.6	21.4	4.8	**638**	3480	**174**	**202**	0.1
	Leal- Taixe' et al. [36]	41.3	55.7	7.1	16.7	640	4776	243	271	8.4
	Pellegrini et al. [46]	32.2	55.7	4.8	**2.4**	1549	4091	893	889	30.6
	MOT baseline [35]	45.4	54.1	16.7	7.1	1407	3593	268	296	83.5
	Wen et al. [63]	47.2	56.6	11.9	4.8	1,140	3,710	245	292	1.9
	Milan et al. [42]	37.5	70.7	4.8	16.7	**638**	5,200	189	209	0.3
Deep Neural Network	Milan et al. [41]	38.3	71.6	9.5	14.3	1,016	4,611	320	417	**165.2**
	Sadeghian et al. [53]	47.0	70.5	11.9	9.5	616	4,236	254	397	1.9
	Bae et al. [4]	42.5	69.3	7.1	7.1	934	4,409	196	438	2.3
	Wang et al. [62]	49.6	70.7	11.9	11.9	780	3,886	192	218	1.7
	Leal-Taix et al. [34]	34.5	69.7	7.1	19.0	672	5,364	282	424	52.8
	Son et al. [58]	49.0	72.6	16.7	7.1	686	3,947	285	380	3.7
	Proposed	**57.6**	**72.8**	**28.6**	4.7	802	**3043**	224	212	78.4

Table 4. 3D MOT 2015 results for AVG-TownCentre. Arrows indicate favourable direction of each metric. Best values are printed in bold

Type	Tracker	MOTA ↑	MOTP ↑	MT (%) ↑	ML (%) ↓	FP ↓	FN ↓	IDS ↓	Frag ↓	Hz ↑
Probabilistic	Klinger et al. [31]	42.4	57.1	**28.8**	20.4	1272	2697	149	173	0.1
	Klinger et al. [32]	42.2	57.4	26.5	19.5	1175	2820	137	184	0.1
	Leal- Taixe' et al. [36]	28.7	51.9	15.0	22.6	1391	3430	277	330	8.4
	Pellegrini et al. [46]	15.2	51.4	7.1	**16.8**	1612	3508	945	797	30.6
	MOT baseline [35]	23.2	52.2	21.7	18.1	2181	3000	312	363	83.5
	Wen et al. [63]	16.8	54.2	11.1	29.2	1,917	3,744	287	319	1.9
	Milan et al. [42]	8.2	69.9	2.7	69.5	763	5,766	**30**	**84**	0.3
Deep Neural Network	Milan et al. [41]	13.4	68.8	3.5	41.2	1,206	4,682	299	414	**165.2**
	Sadeghian et al. [53]	36.2	69.5	26.1	17.7	1,448	2,882	234	389	1.9
	Bae et al. [4]	30.7	68.9	13.7	31.9	1,013	3,807	136	367	2.3
	Wang et al. [62]	31.3	69.5	16.8	33.2	952	3,825	137	246	1.7
	Leal-Taix et al. [34]	19.3	69.0	4.4	44.7	**698**	4,927	142	289	52.8
	Son et al. [58]	30.8	**69.8**	18.1	31.4	1,191	3,643	111	409	3.7
	Proposed	**42.5**	**69.8**	27.0	19.5	1182	2826	139	186	78.4

tracker reflects our observations.

Son et al. [58] associates the temporal distance between frame patches in a video with appearance based feature matching using CNNs. They utilise detection properties such as centre position, height, velocity and temporal distance in addition to the convolution features, in their data association framework. Yet their method leads to very high fragmentation of trajectories, frequent id switches and higher FN values while also having a high time complexity (i.e a lower Hz) in both sequences.

Furthermore, we would like to draw comparisons with [41], which utilises a deep learning based trajectory planing method for data association and eschews appearance features. Even though with their single LSTM based approach, which does not incorporate any neighbourhood or context information, they were able to obtain a low computational complexity, the method leads to frequent id swapping (i.e higher IDS), higher fragmentation and a very low MT value. Our proposed approach, while similar in that we rely on motion prediction only, has an increased capacity through incorporating the complete history of the neighbourhood as well as the contextual factors derived directly from trajectory modelling. Thus the proposed model has been able to generate results with higher performance.

Impact of using multiple trajectory predictions. We investigate the contributions of each prediction component in our framework for the data association task by measuring the performance of using each component separately and together in terms of tracking performance, in the training set of 3D MOT 2015 benchmark. A detailed evaluation of each tracking method against various MOT matrices is presented in Tab. 5. The details on each system are as follows:

T.1 System with only data association with short term trajectory prediction (STP) (i.e Eq. 12)

T.2 System with STP and trajectory update with only spatial dissimilarity (i.e Eq. 12 + Eq. 13)

T.3 System with STP and trajectory update with only context dissimilarity (i.e Eq. 12 + Eq. 14)

T.4 System with STP and trajectory update with combined dissimilarity (i.e Eq. 12 + Eq. 13 + Eq. 14)

The predicted short term trajectory is the central driving mechanism in the proposed framework, due to the fact that most interactions and occlusions occur over short periods of time. It should be pointed out that each component (SD

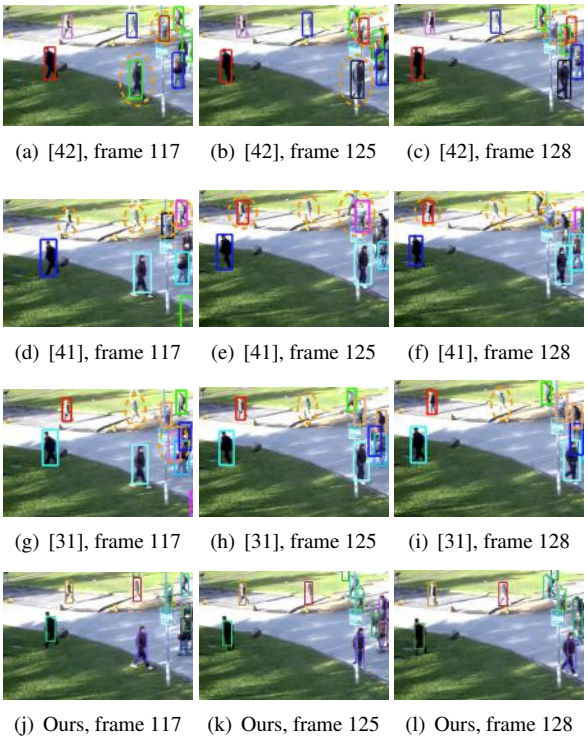
Table 5. Analysis of the contribution of each component of the proposed tracking framework on the training set of the MOT benchmark. Arrows indicate favourable direction of each metric. Best values are printed in bold

Tracker	MOTA ↑	MOTP ↑	MT (%) ↑	ML (%) ↓	FP ↓	FN ↓	IDS ↓	Frag ↓
T1	42.3	48.2	26.7	5.5	1032	2687	329	330
T2	51.4	56.3	27.3	4.5	877	2520	207	184
T3	51.4	**65.9**	28.2	**3.7**	892	2430	177	191
T4	**57.9**	**65.9**	**31.5**	3.6	**506**	**2187**	**146**	**105**

Table 6. Comparison of results from the proposed approach against state-of-the-art methods on ETHMS dataset. Arrows indicate favourable direction of each metric. Best values are printed in bold

Tracker	Recall ↑	Precision ↑	MT (%) ↑	ML (%) ↓	Frag ↓	IDs ↓
DP [49]	67.4	**91.4**	50.2	9.9	143	4
PIRMPT [33]	76.8	86.6	58.4	8.0	**23**	11
Online CRF [66]	79.0	90.4	68.0	7.2	19	11
DCEM [42]	76.2	87.6	58.3	7.1	78	43
Proposed	**89.8**	91.0	**73.8**	**7.3**	25	**3**

and CD) of the proposed trajectory update mechanism positively impacts on the overall performance as it lowers the possibility of ID switches and trajectory fragmentation. The results on the trajectory update process with the combination of the dissimilarity measures implies that exploiting the data association with the proposed update process can significantly improve the performance of the tracker.



(a) [42], frame 117 (b) [42], frame 125 (c) [42], frame 128

(d) [41], frame 117 (e) [41], frame 125 (f) [41], frame 128

(g) [31], frame 117 (h) [31], frame 125 (i) [31], frame 128

(j) Ours, frame 117 (k) Ours, frame 125 (l) Ours, frame 128

Figure 3. Qualitative evaluation: Results on 3dMOT 2015 PETS09-S2L2 sequence. Row 1-3 shows the tracking results of Milan et al. [42], Milan et al. [41] and Klinger et al. [31] respectively whereas the fourth row presents the results of the proposed method. The dashed circles indicates the ID switches and missed pedestrians during the tracking

A qualitative evaluation of the tracking results is shown in Fig. 3, where we compare the tracking outputs from the proposed method with state-of-the-art baselines. The ID switches and tracker misses are highlighted in dashed circles. It can be observed that Milan et al. [41] performs poorly with frequent ID switches and missed tracks, as the tracker lacks neighbourhood information. The temporal modelling with probabilistic networks methods, Milan et al. [42] and Klinger et al. [31], fail to anticipate the motion and generate erroneous tracking results. In contrast, the proposed framework utilises the context and neighbourhood dynamics and generates accurate tracking results, even without using an appearance model.

We also evaluate the proposed tracking framework on the ETHMS dataset (see Tab. 6) where a busy pedestrian street is filmed from a moving camera. We use pedestrian detections from the proposed detector and publicly available evaluation script. Methods like DP [49], Online CRF [66] and PIRMPT [33] are highly reliant on appearance based tracklet linking and occlusion avoidance. Yet our efficient data association scheme based on trajectory prediction outperforms these state-of-the-art methods with fewer ID switches and higher precision and recall. The DCEM [42] approach replaces appearance based data association with trajectory modelling, yet fails to generate accurate tracking results compared to the proposed method.

5. Conclusion

This paper has presented a new online deep learning framework for muli-person localisation and tracking. The proposed localisation framework builds upon generative adversarial networks and performs sequential modelling, allowing us to localise pedestrians in cluttered environments even in the presence of noise and other image artefacts. Our data association scheme utilises a trajectory modelling approach and anticipates human behavioural patterns under different contexts. It not only results in a light weight framework compared to other CNN based person re-identification architectures, but also introduces intelligence into the tracking framework via modelling different human behavioural patterns under different contexts such as group motion and random exploration. Our evaluations on publicly available benchmarks have shown that the proposed method exhibits superior performance, especially among current state-of-the-art deep learning methods. The evaluations on both static and dynamic cameras ensures the applicability of the proposed method in variety of applications including autonomous driving, robotics, and egocentric vision.

Acknowledgement

This research was supported by the Australian Research Council's Linkage Project LP140100282 "Improving Productivity and Efficiency of Australian Airports". The authors also thank QUT High Performance Comput-

ing (HPC) for providing the computational resources for this research.

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[2] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1265–1272. IEEE, 2011.

[3] M. Babaee, Y. You, and G. Rigoll. Combined segmentation, reconstruction, and tracking of multiple targets in multi-view video sequences. *Computer Vision and Image Understanding*, 154:166–181, 2017.

[4] S.-H. Bae and K.-J. Yoon. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[5] P. Baqué, F. Fleuret, and P. Fua. Deep occlusion reasoning for multi-camera multi-target detection. *arXiv preprint arXiv:1704.05775*, 2017.

[6] J. Black, T. Ellis, and P. Rosin. Multi view image surveillance and tracking. In *Motion and Video Computing, 2002. Proceedings. Workshop on*, pages 169–174. IEEE, 2002.

[7] A. A. Butt and R. T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1846–1853, 2013.

[8] T. Chavdarova and F. Fleuret. Deep multi-camera people detection. *arXiv preprint arXiv:1702.04593*, 2017.

[9] Y. Dorai, F. Chausse, S. Gazzah, and N. E. B. Amara. Multi target tracking by linking tracklets with a convolutional neural network. In *VISIGRAPP (6: VISAPP)*, pages 492–498, 2017.

[10] A. Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[11] A. Ess, B. Leibe, K. Schindler, and L. Van Gool. A mobile vision system for robust multi-person tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[12] A. Ess, K. Schindler, B. Leibe, and L. Van Gool. Improved multi-person tracking with active occlusion handling. In *ICRA Workshop on People Detection and Tracking*, volume 2, 2009.

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[14] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes. Tree memory networks for modelling long-term temporal dependencies. *arXiv preprint arXiv:1703.04706*, 2017.

[15] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Going deeper: Autonomous steering with neural memory networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *arXiv preprint arXiv:1702.05552*, 2017.

[17] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Task specific visual saliency prediction with memory augmented conditional generative adversarial networks. *Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on*, 2018.

[18] T. Fernando, X. Wei, C. Fookes, S. Sridharan, and P. Lucey. Discovering methods of scoring in soccer using tracking data. *KDD Workshop on Large Scale Sports Analytics*, 2015.

[19] J. Ferryman and A. Shahrokni. Pets2009: Dataset and challenge. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1–6. IEEE, 2009.

[20] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes. Two stream lstm: A deep fusion framework for human action recognition. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 177–186. IEEE, 2017.

[21] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *European Conference on Computer Vision*, pages 324–337. Springer, 2010.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[23] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[24] D. Hoiem, A. A. Efros, and M. Hebert. Putting objects in perspective. *International Journal of Computer Vision*, 80(1):3–15, 2008.

[25] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4073–4082, 2015.

[26] X. Jiang, Y. Pang, X. Li, and J. Pan. Speed up deep neural network based pedestrian detection by sharing features across multi-scale models. *Neurocomputing*, 185:163–170, 2016.

[27] I. N. Junejo, O. Javed, and M. Shah. Multi feature path modeling for video surveillance. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 716–719. IEEE, 2004.

[28] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):319–336, 2009.

[29] S. Kim, S. Kwak, J. Feyereisl, and B. Han. Online multi-target tracking by large margin structured learning. In *Asian*

Conference on Computer Vision, pages 98–111. Springer, 2012.

[30] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] T. Klinger, F. Rottensteiner, and C. Heipke. Probabilistic multi-person tracking using dynamic bayes networks. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2015.

[32] T. Klinger, F. Rottensteiner, and C. Heipke. Probabilistic multi-person localisation and tracking in image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 127:73–88, 2017.

[33] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1217–1224. IEEE, 2011.

[34] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2016.

[35] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015.

[36] 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.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[38] J. Lou, Q. Liu, T. Tan, and W. Hu. Semantic interpretation of object activities in a surveillance system. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 777–780. IEEE, 2002.

[39] C.-W. Lu, C.-Y. Lin, C.-Y. Hsu, M.-F. Weng, L.-W. Kang, and H.-Y. M. Liao. Identification and tracking of players in sport videos. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 113–116. ACM, 2013.

[40] B. Majecka. Statistical models of pedestrian behaviour in the forum. *MSc Dissertation, School of Informatics, University of Edinburgh*, 2009.

[41] A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, pages 4225–4232, 2017.

[42] A. Milan, K. Schindler, and S. Roth. Multi-target tracking by discrete-continuous energy minimization. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2054–2068, 2016.

[43] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *European Conference on Computer Vision*, pages 28–39. Springer, 2004.

[44] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3258–3265. IEEE, 2012.

[45] W. Ouyang and X. Wang. Single-pedestrian detection aided by multi-pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3198–3205, 2013.

[46] 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.

[47] P. Peng, Y. Tian, Y. Wang, J. Li, and T. Huang. Robust multiple cameras pedestrian detection with multi-view bayesian network. *Pattern Recognition*, 48(5):1760–1772, 2015.

[48] A. Petrovskaya and S. Thrun. Model based vehicle detection and tracking for autonomous urban driving. *Autonomous Robots*, 26(2-3):123–139, 2009.

[49] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1201–1208. IEEE, 2011.

[50] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[51] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. *arXiv preprint arXiv:1612.08242*, 2016.

[52] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[53] 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.

[54] A. Sanborn and J. Skryzalin. Deep learning for semantic similarity. *CS224d: Deep Learning for Natural Language Processing. Stanford, CA, USA: Stanford University*, 2015.

[55] K. Schindler, A. Ess, B. Leibe, and L. Van Gool. Automatic detection and tracking of pedestrians from a moving stereo rig. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6):523–537, 2010.

[56] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633, 2013.

[57] L. Shafarenko, M. Petrou, and J. Kittler. Automatic watershed segmentation of randomly textured color images. *IEEE transactions on Image Processing*, 6(11):1530–1544, 1997.

[58] J. Son, M. Baek, M. Cho, and B. Han. Multi-object tracking with quadruplet convolutional neural networks. *CVPR*, 2017.

[59] X. Song, J. Cui, H. Zha, and H. Zhao. Vision-based multiple interacting targets tracking via on-line supervised learning. In *European Conference on Computer Vision*, pages 642–655. Springer, 2008.

[60] Y.-m. Song and M. Jeon. Online multiple object tracking with the hierarchically adopted gm-phd filter using motion

and appearance. In *Consumer Electronics-Asia (ICCE-Asia), IEEE International Conference on*, pages 1–4. IEEE, 2016.

[61] S. Tian, F. Yuan, and G.-S. Xia. Multi-object tracking with inter-feedback between detection and tracking. *Neurocomputing*, 171:768–780, 2016.

[62] B. Wang, L. Wang, B. Shuai, Z. Zuo, T. Liu, K. Luk Chan, and G. Wang. Joint learning of convolutional neural networks and temporally constrained metrics for tracklet association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

[63] L. Wen, Z. Lei, M.-C. Chang, H. Qi, and S. Lyu. Multi-camera multi-target tracking with space-time-view hyper-graph. *International Journal of Computer Vision*, 122(2):313–333, 2017.

[64] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.

[65] J. Xing, H. Ai, L. Liu, and S. Lao. Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling. *IEEE Transactions on Image Processing*, 20(6):1652–1667, 2011.

[66] B. Yang and R. Nevatia. An online learned crf model for multi-target tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2034–2041. IEEE, 2012.

[67] Z. Zhang, K. Huang, and T. Tan. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1135–1138. IEEE, 2006.