Bifold and Semantic Reasoning for Pedestrian Behavior Prediction

Amir Rasouli, Mohsen Rohani, Jun Luo
Huawei Technologies Canada
{amir.rasouli, mohsen.rohani, jun.luo1}@huawei.com

Abstract

Pedestrian behavior prediction is one of the major challenges for intelligent driving systems. Pedestrians often exhibit complex behaviors influenced by various contextual elements. To address this problem, we propose BiPed, a multitask learning framework that simultaneously predicts trajectories and actions of pedestrians by relying on multimodal data. Our method benefits from 1) a bifold encoding approach where different data modalities are processed independently allowing them to develop their own representations, and jointly to produce a representation for all modalities using shared parameters; 2) a novel interaction modeling technique that relies on categorical semantic parsing of the scenes to capture interactions between target pedestrians and their surroundings; and 3) a bifold prediction mechanism that uses both independent and shared decoding of multimodal representations. Using public pedestrian behavior benchmark datasets for driving, PIE and JAAD, we highlight the benefits of the proposed method for behavior prediction and show that our model achieves state-of-the-art performance and improves trajectory and action prediction by up to 22% and 9% respectively. We further investigate the contributions of the proposed reasoning techniques via extensive ablation studies.

1. Introduction

Predicting road user behavior in complex urban environments is fundamental for assistive and intelligent driving systems. Prediction is particularly challenging when these systems are encountering pedestrians who exhibit diverse behaviors [1] that depend on various contextual factors, such as social interactions, road structure, traffic condition, and other environmental factors [2].

Pedestrian behavior can be predicted implicitly in the form of future trajectories [3, 4, 5], or explicitly in the form of upcoming actions [6, 7, 8]. It is evident from recent studies [5, 9, 10] that both types of behavior prediction play complementary roles. For instance, predicting pedestrian actions, such as crossing the road, implies the possibility of a lateral motion across the road. Similarly, a pedestrian approaching a parked vehicle is expected to interact with it. To capture these complementary aspects of behavior, we propose a multitask learning framework that simultaneously predicts trajectories, actions, and final locations of pedestrians (see Figure 1). To learn complex pedestrian behavior, our model relies on multiple data modalities including visual context, pedestrian motion, and ego-vehicle dynamics. The proposed method independently and jointly processes different input modalities and tasks. Independent processing allows each modality or task to learn its own parameters, whereas joint processing acts as a regularizer, inducing the model to learn more representative features. Since pedestrians' behaviors are often influenced by what is around them, we introduce a novel technique to model interactions between target pedestrians and their surroundings based on the semantic composition of the scenes. The proposed technique relies on visuospatial semantic representations of the scenes divided into categories based on object classes.

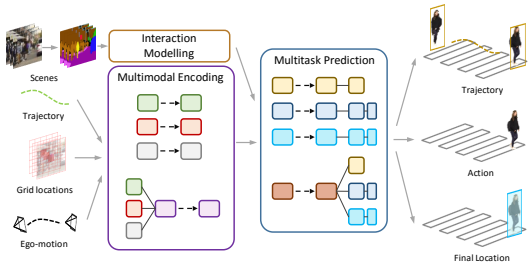

Figure 1. Proposed multitask learning for simultaneous prediction of pedestrian trajectory, action and final location. Interaction of traffic elements are modeled along with separately and jointly encoded visual context, pedestrian motion, and ego-motion inputs.

We evaluate the performance of the proposed method using public pedestrian behavior benchmark datasets, PIE [10] and JAAD [11], and show that our method significantly improves over state-of-the-art algorithms on both pedestrian trajectory and action prediction tasks.

2. Related Works

2.1. Multimodal Behavior Prediction

Human behavior prediction research in computer vision and robotics has many practical applications, such as

human-object [12, 13] and human-human [14, 15] interaction, risk assessment [16, 17], anomaly detection [18], surveillance [19, 9], sports forecasting [20, 21], and intelligent driving systems [22, 8]. As for pedestrians, their behaviors can be predicted implicitly in terms of trajectories [23, 24, 3, 25, 26] or explicitly in terms of actions, such as crossing the road [27, 8], or interacting with objects [9].

Trajectory Prediction. Pedestrian prediction research is dominated by trajectory prediction. A large body of work in this domain is dedicated to prediction on surveillance sequences where the movements of groups of pedestrians are observed from a fixed bird's eye view perspective [23, 24, 3, 25, 26, 28, 29, 30, 31].

Opposed to bird's eye view prediction algorithms are ego-centric methods that predict the trajectories of pedestrians in the image plane recorded from the perspective of a moving camera [4, 5, 10, 32, 33, 34, 35, 36]. Ego-centric trajectory prediction is generally more challenging because, first, in the absence of depth information, the relative positions of agents are difficult to infer and, second, the ego-motion of the camera can influence both the behavior of pedestrians and the predicted trajectories in the image plane. To address these challenges, ego-centric algorithms use multimodal approaches. For example, the model in [35] proposes a two-stream recurrent encoder-decoder architecture where one stream processes pedestrian trajectories and the other ego-motion of the vehicle. The output of the ego-motion predictor in conjunction with the pedestrian stream is used to forecast trajectories. Some models based on a similar architecture use pedestrian action (e.g. waiting to cross [5]) or intention of performing an action (e.g. crossing the road [10]) as inputs to infer future trajectories. Some methods rely on fully feedforward approaches, such as [32], where three streams of 1D convolutional layers encode ego-motion, pedestrians' trajectories, and poses, the output of which are decoded using a convolutional decoder.

Action Prediction. Pedestrian action prediction is also being actively investigated with an emphasis on prediction of interactions between humans or groups [37, 15, 38, 39, 40, 41]. In the driving context, the main focus is on predicting pedestrian crossing action to assess the risks and anticipate pedestrian trajectories for safe motion planning [42, 6, 43, 44, 11, 7, 8, 27]. A group of these algorithms use unimodal image sequences for prediction. For example, the methods of [44, 6] are convolutional generative models that predict future traffic scenes, which are then used to predict crossing events. In [43] a 3D DenseNet is used to first localize pedestrians and then predict their crossing actions.

For crossing prediction, multimodal architectures are more common [11, 7, 8, 27]. For instance, the model in [8] is a multi-level recurrent architecture, which receives as input the appearance of pedestrians and their surrounding context, pedestrians' trajectories and poses as well as the ego-vehicle speed. The features are gradually fed into the network at each level according to their complexity. The method proposed in [27] uses a hierarchical LSTM architecture in which visual features, including optical flow maps and images, and vehicle dynamics are encoded using independent LSTMs. The outputs of these LSTMs are concatenated and fed into an embedding layer followed by another LSTM prior to prediction. The method in [42] is a hybrid architecture that encodes visual features using a 3D convolution network and other modalities using LSTMs followed by temporal attention modules. The representations are fed into a modality attention unit prior to prediction.

Independent encoding of different data modalities, as proposed previously, however, is more susceptible to noise and may not capture cross-modal correlations. Hence, we propose a bifold approach that encodes different modalities both independently and jointly, thus inducing the system to learn more representative features for different tasks, while learning temporal correlations between different modalities.

2.2. Interaction Modeling

One of the fundamental components of behavior prediction in a multi-agent setting is the ability to understand the interactions between agents as their behaviors can potentially impact one another. Thus, interaction modeling is widely used for pedestrian trajectory prediction. For instance, the methods in [31, 25, 26] use the social pooling technique, which jointly processes trajectories of pedestrians within a neighboring region to learn the spatial dependencies among them. Attention-based approaches assign importance values to interacting pedestrians according to their relative distance and motion [30, 29, 45]. Alternatively, graph structures can be used to assign importance to interacting pedestrians represented as nodes [46, 24, 3, 47, 48].

Ego-centric trajectory prediction algorithms follow a similar trajectory-based route to model interactions. For example, the method in [36] defines two regions for modeling interactions between traffic participants: a region around the ego-vehicle defined by an ellipsoid on the image plane and a driving horizon of the vehicle that is determined based on the actions of the driver. The authors of [5] model the interactions by jointly processing the past locations and action information of all agents in the scene using an RNN.

In the absence of global positions of objects, modeling interactions based on 2D coordinates in the image plane can be problematic. For example, two people of different heights walking next to each other can have trajectories similar to those of two people of the same height but far apart from each other. In addition, pair-wise modeling of interactions between each pedestrian and all other road users is not scalable, especially in urban driving scenarios where many traffic participants potentially interact with each other. To

address these shortcomings, we propose a categorical interaction modeling technique, which relies on the visuospatial changes of different categories of objects over time. Our method encodes both dynamic and static context of the scenes and uses an attention mechanism to determine the importance of different contextual elements.

2.3. Multitask Learning

As evidenced by various domains of machine learning, such as action/expression understanding [49, 50, 51, 52], object recognition [53, 54, 55, 56], intelligent driving [57, 58, 59, 60, 61], and other computer vision applications [62, 63, 64, 65], multitask learning is an effective way for improving the performance on multiple tasks.

Pedestrian behavior prediction is not an exception and in recent years a number of multitask algorithms have been proposed to solve this problem [66, 67, 9, 68]. For example, the authors of [66] simultaneously predict pedestrian head poses and trajectories and exploit the correlation between the two to improve trajectory prediction. The method in [68] detects pedestrians and predicts their future trajectories at the same time . It uses point cloud sequences encoded into different feature representations which are fed into a feedforward backbone network. The output of the backbone is used to generate temporal proposals for localizing pedestrians and predicting their trajectories. Liang et al. [9], jointly predict pedestrian trajectory and activity, e.g. interacting with a car, using a recurrent framework that encodes different data types, including scene semantics, poses, and trajectories, and combines their representations for joined reasoning on different tasks in separate prediction branches.

In the context of intelligent driving, predicting both trajectories and actions of pedestrians is important for planning. These tasks can play complementary roles – trajectory prediction provides accurate location information in future frames, whereas action prediction helps interpret the nature of events and the types of motions to be expected. Given such a mutually beneficial relationship between these two tasks, we propose a multitask learning approach that predicts trajectories and actions simultaneously. Unlike other approaches, in addition to independent task predictors, we use a shared prediction module to capture correlation across different tasks, and make final predictions based on the outputs of both independent and joint modules.

Contributions: 1) We propose a multitask pedestrian behavior prediction framework to simultaneously predict pedestrian trajectory, action and final location in urban traffic scenes using ego-centric image sequences. The proposed approach benefits from a bifold encoding scheme that independently and jointly learns different data modalities, a categorical interaction module which encodes the interactions between target pedestrians and their surroundings, and a prediction mechanism which uses independent and shared decoders. 2) Using two publicly available pedestrian behavior datasets, namely PIE [10] and JAAD [11], we evaluate our model, which achieves state-of-the-art performance on both tasks of trajectory and crossing action prediction. 3) We show the advantages of the components of the proposed model by conducting extensive ablation studies.

3. Method

3.1. Problem Formulation

We formulate pedestrian behavior prediction as a multi-objective optimization process in which the goal is to learn distribution $p(L_p, a_i, g_i^{t+\tau} | SC_o, L_o, G_o, V)$ for some pedestrian $1 < i < n$ where $L_p = \{l_i^{t+1}, l_i^{t+2}, ..., l_i^{t+\tau}\}$, $a_i \in \{0, 1\}$, and $g_i^{t+\tau} \in \{0, 1, ..., k\}$ are future trajectory, crossing action, and the final location of pedestrian i in a grid on the image plane. Predictions are based on observed scenes $SC_o = \{sc^{t-m+1}, sc^{t-m+2}, ..., sc^t\}$, the pedestrian's trajectory $L_o = \{l_i^{t-m+1}, l_i^{t-m+2}, ..., l_i^t\}$, their grid locations $G_o = \{g_i^{t-m+1}, g_i^{t-m+2}, ..., g_i^t\}$ and the ego-vehicle motion $V = \{v^{t-m+1}, v^{t-m+2}, ..., v^{t+\tau}\}$. Here, m denotes observation duration and τ is prediction duration.

3.2. Architecture

Our approach simultaneously predicts future trajectories, actions and final locations of pedestrians (see Figure 2). Below we discuss different modules of the proposed model:

Context encoding deals with processing and encoding of multimodal observation inputs.

Categorical interaction module models the interactions between target pedestrians and surrounding traffic elements.

Behavior prediction outputs future trajectories of pedestrians, the probabilities of their crossing actions, and final locations in the image grid map.

3.3. Context Encoding

We use four different input modalities: scene images, pedestrians' trajectories, their grid locations, and ego-vehicle motion, in order to encode context.

Scenes are RGB images of traffic scenes recorded from an ego-centric perspective capturing the view in front of the ego-vehicle. These images capture visible changes in the scene. The semantic segmentation module processes scene images and generates semantic maps of the traffic elements $sm^{t-m+1:t}$. These maps are fed into Categorical Interaction Module (CIM) (see Section 3.4) to generate a representation C_{int} that encodes the interactions between target pedestrians and traffic elements.

Pedestrian trajectory. In the context of prediction in the 2D image plane, in addition to estimating future trajectories, it is also important to localize pedestrian boundaries. As suggested in the past works [35, 10], predicting bounding box coordinates, as opposed to center coordinates, can

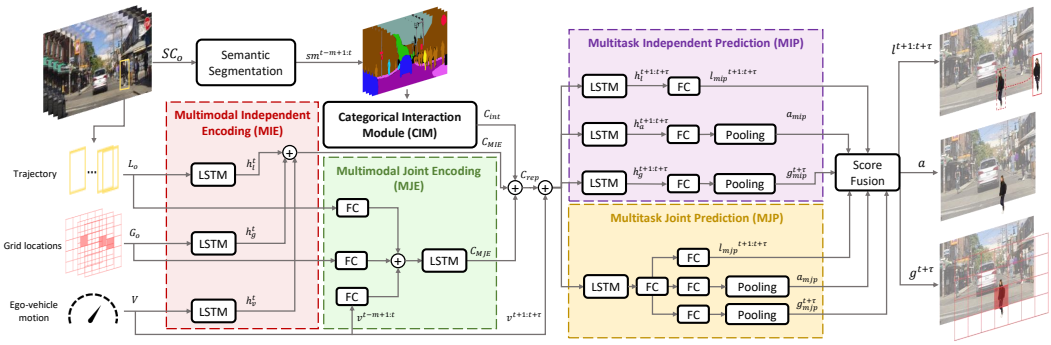

Figure 2. The diagram of the proposed multitask pedestrian behavior prediction method. Our model relies on 4 input modalities, namely scene images, and dynamics, including pedestrians' trajectories, their grid locations, and the ego-vehicle motion. The dynamics are encoded both by Multimodal Independent Encoding (MIE) and Multimodal Joint Encoding (MJE) modules. The scene images are converted to semantic maps and fed into Categorical Interaction Module (CIM) to generate interaction representation C_{int}. The dynamics encodings are concatenated with C_{int} to form context representation C_{rep}, which is combined with planned ego-motion and fed into prediction modules, Multitask Independent Prediction (MIP) and Multitask Joint Prediction (MJP), the outputs of which are averaged for final predictions of pedestrian trajectory ($l^{t+1:t+\tau}$), action (a), and final grid location ($g^{t+\tau}$).

improve trajectory prediction because bounding box coordinates implicitly capture the changes in relative distance between the ego-vehicle and pedestrians as well as the camera ego-motion. Hence, we use spatial coordinates of bounding boxes around pedestrians defined by top-left and bottom-right corner points $[(x_1, y_1), (x_2, y_2)]$ to encode the changes in the locations of the pedestrians.

Grid locations. Inspired by [9], we convert the location information of pedestrians into grid classes. To achieve this, the image plane is divided into $N \times M$ grid cells each of which is assigned with a unique class. We identify the corresponding grid of pedestrian i at time t by $g^t = \text{argmin}_{j \in cls}(|center_{g_j} - center_{l_i^t}|)$ or the cell whose center is closest to the center of the pedestrian's bounding box. Here, cls refers to classes associated to grid locations.

Ego-vehicle motion reflects the changes in the state of the ego-vehicle over time denoted as $v^t = [s^t, v_x^t, v_z^t]$ where s is the speed of the vehicle and v_x and v_z are velocities of the vehicle along x and z axes.

3.3.1 Bifold context representation

In behavior prediction domain, it is a common practice to process different modalities separately in a unimodal setting, meaning that, first, a separate feature representation is generated for each modality and then these representations are fused prior to inference [35, 9, 8]. For example, when using recurrent networks, the last hidden layers, h^t, of networks are concatenated. This approach allows each modality to be learned with its own parameters without the noise introduced by other modalities. Independent processing, however, does not capture cross-modal correlations in the temporal dimension and is also potentially susceptible to missing data or noise [69]. An alternative is hard parameter sharing, where all modalities are jointly learned using a single model. Parameter sharing can act as a regularizer in a multitask learning framework inducing the model to learn more representative features. To benefit from both approaches, we employ a bifold mechanism to encode input data, namely trajectories, grid locations, and vehicle states.

Multimodal Independent Encoding (MIE). This module generates an independent representation for each modality. Each data input is fed into a Recurrent Neural Network (RNN). The last hidden states of the RNNs are concatenated to form a unified representation, $C_{MIE} = h_l^t \oplus h_g^t \oplus h_v^t$, where $\oplus$ is the concatenation operation and l, g, and v stand for location, grid and vehicle respectively.

Multimodal Joint Encoding (MJE). This module jointly encodes different data modalities. Here, it is necessary to project the data from different modalities into a common feature space. MJE generates C_{MJE} by applying an embedding layer to each input modality, and then concatenating the outputs of embedding layers and processing the concatenated representation using a single RNN.

The final context representation is generated by concatenating all three contextual representations as,

$$C_{rep} = C_{int} \oplus C_{MIE} \oplus C_{MJE} \quad (1)$$

3.4. Categorical Interaction Module (CIM)

As discussed earlier, in an ego-centric setting without depth information the trajectories of pedestrians in the image plane are not sufficient, and perhaps misleading, for modeling interactions between different agents. To remedy this issue, we rely on semantic parsing of the scenes and implicitly model interactions between target pedestrians and different groups of traffic elements (see Figure 3). We first generate semantic maps using the input scene images to identify the position and category of each object. Then, the maps are divided into different categories, namely the target pedestrian (p), people surrounding the pedestrian (pl), motorcyclists/bicyclists (b), vehicles (v) (e.g. cars, buses, trucks) and static context (st) (e.g. signs, roads, signals).

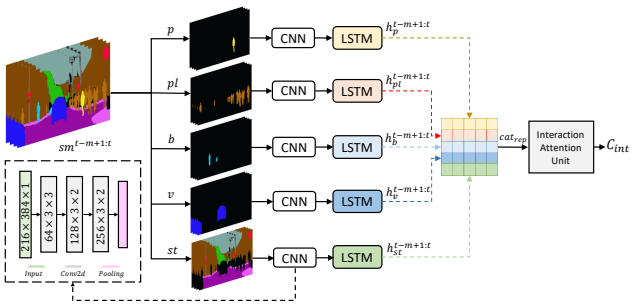

Figure 3. Categorical interaction module. Semantic maps are divided into the target pedestrian (p), people surrounding the pedestrian (pl), motorcyclists/bicyclists (b), vehicles (v) and static context (st). The maps in each category are processed using convolutional layers followed by an LSTM to generate spatio-temporal representations, which are concatenated and fed into the Interaction Attention Unit (IAU) to produce a weighted representation.

Semantic categories are processed using multiple 2D convolution layers followed by recurrent networks. The hidden states of the RNNs are concatenated to form a shared categorical representation as follow:

$$cat_{rep} = h^o_p \oplus h^o_{pl} \oplus h^o_b \oplus h^o_v \oplus h^o_{st} \in \mathbb{R}^{m \times f} \quad (2)$$

where observation length o is m and f is the total size of hidden units in recurrent networks. The shared representation is fed into Interaction Attention Unit (IAU) to generate categorical interaction context C_{int}.

Interaction Attention Unit (IAU). Inspired by [70], IAU is an attention method that receives as input temporal data and outputs a unified weighted representation. Denoting h^t as the last time step of the input sequence to the attention unit, we first generate attention scores by measuring the similarity between the last and every other time steps,

$$s^i = h^{t\prime} W_a h^i \quad (3)$$

where $'$ is the transpose operation. Using the scores, attention weights per time step are computed by $\alpha^i = \mathrm{softmax}(s^i)$ and used to calculate the context vector as

$$c^t = \sum_{i \in [t-m+1:t]} \alpha^i h^i. \quad (4)$$

A combination of the context vector and last time step representation is used to generate interaction context,

$$C_{int} = \tanh(W_c[c^t \oplus h^t]) \quad (5)$$

where $C_{int} \in \mathbb{R}^{1 \times q}$ and $\oplus$ is the concatenation operation.

3.5. Behavior Prediction

Predictions are made based on the concatenation of context representation, C_{rep} with future ego-vehicle motion $v^{t+1:t+\tau}$. As in the encoding step, we use a bifold mechanism with two modules: Multitask Independent Prediction (MIP) and Multitask Joint Prediction (MJP).

Multitask Independent Prediction (MIP). A separate recurrent decoder branch is used for each task to produce three predictions: $l^{t+1:t+\tau}_{mip}$, a_{mip} and $g^{t+\tau}_{mip}$ that represent future trajectories, action and final location of the pedestrian on the grid respectively.

Trajectories are 2D bounding boxes defined by $[(x_1, y_1), (x_2, y_2)]$ denoting the top-left and bottom-right corners of each box. Predictions are made by a linear transformation of hidden states of the trajectory branch.

Action. Since the focus of this paper is on intelligent driving systems, we predict pedestrian crossing action, i.e. at a given time, we predict whether the pedestrian will cross in front of the ego-vehicle. To estimate the probability of pedestrian crossing at each time step we perform a linear transformation followed by a sigmoid activation. Then, a global average pooling is used to calculate the mean of predictions over all time steps as the final prediction probability. More formally, the future action is given by,

$$a_{mip} = \frac{\sum_{i \in [t+1:t+\tau]} \sigma(f(h^i))}{\tau}. \quad (6)$$

Final grid location prediction is an auxiliary task. As argued in [9], grid location can act as a bridge between trajectory and action prediction and perform as a regularizer by indicating a final destination for the predicted trajectory and the possibility of an action, in this case crossing, i.e. whether a pedestrian's final location falls on the road in the path of the ego-vehicle. Unlike the previous approach, for computational efficiency, we only use a single scale grid set based on the range of pedestrians' scales in the image plane (see Section 4.6.2 for an ablation study). We treat the grid prediction task as a classification problem and predict the class of the final grid location on the image plane. Following the same procedure as action prediction task, the final grid location at each time step is given by $\mathrm{softmax}(f(h^t))$ and averaged over all time steps.

Multitask Joint Prediction (MJP). Unlike MIP, this module uses a single RNN as a shared decoder, the output of which is processed using a fully connected (fc) layer followed by three separate branches for each task. The predictions are made the same way as in MIP.

Score fusion. In order to calculate the final prediction scores for each task, we follow the approach in two-stream methods as in [71], and compute the final scores as follows: $l^k = \frac{1}{N}\sum_i l^k_i$ where $k = t+1, ..., t+\tau$ and $f = \frac{1}{N}\sum_i f_i$, where $f \in \{a, g^{t+\tau}\}$, $N = 2$ and $i \in \{mip, mjp\}$.

3.6. Learning Objectives

The model is trained end-to-end using a multi-objective loss function. For trajectory prediction we use,

$$L_l = \sum_{i=1}^{n} \sum_{j=t}^{t+\tau} \log(\cosh(y^j_i - \hat{y}^j_i)), \quad (7)$$

which compared to commonly used L2 loss as in [9, 10], is less prone to outliers and generally converges faster. For action prediction, we use a binary cross-entropy loss,

$$L_a = -\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \quad (8)$$

and for grid location prediction, use a multi-class entropy,

$$L_g = -\sum_{i=1}^{i} \sum_{gc} y_{i,gc} \log(\hat{y}_{i,gc}) \quad (9)$$

where n is the number of samples and gc is the number of grid classes. The final loss is given by,

$$L = \alpha L_l + \beta L_a + \gamma L_g \quad (10)$$

where α , β and γ are loss weights determined empirically.

4. Evaluation

4.1. Implementation

We use LSTMs for all encoders and decoders with hidden size of 256, L2 regularization of 0.0001 and $\mathrm{softsign}$ activations with the exception of trajectory decoder for which $\tanh$ is used. The sizes of the embedding layers in MJE and MJP are set to 64 and 128 respectively. For grid classes, the image is divided into a grid of 18×32 cells of size 60×60 pixels based on the lower bound of pedestrian scales in the image plane within the observation horizon.

For semantic maps, we used the method in [72] pretrained on the CityScapes dataset [73]. The maps are downsampled by a factor of 5 to 384×216 pixels while maintaining the aspect ratio. We use three 2D convolutional layers (see Figure 3 for details) with shared weights across different categories. To encode categories, LSTMs with 128 cells with $\tanh$ activation and $L2$ regularization of 0.0001 are used. The output dimension of the IAU is set to 128.

4.2. Datasets

As the main dataset for our evaluations, we use the **Pedestrian Intention Estimation (PIE)** dataset [10] which consists of 6 hours of driving footage in urban environments. The dataset provides bounding box annotations for pedestrians and traffic objects as well as sensor readings of the ego-vehicle and ego-motion data recorded from the camera. We use the default data split ratios as in [10].

To accommodate both trajectory and action prediction tasks, we clip the pedestrian tracks up to the crossing event frames and sample sequences with 50% overlap and time to event between 1 to 2 seconds (30 to 60 frames) as discussed in [8]. Overall, there are 3980 training sequences out of which 995 are crossing and the rest are non-crossing events.

We also report on the **Joint Attention in Autonomous Driving (JAAD)** dataset [11], which contains short clips of urban driving scenes. Compared to PIE, the JAAD dataset is less diverse, has shorter sequences and fewer crossing samples, and does not contain ego-motion data. Instead, JAAD has high-level driver actions, e.g. *moving slow*, *speeding up*, describing the state of the ego-vehicle. We use this information in place of ego-motion data and split the data similar to [74]. Training samples are generated similar to PIE resulting in 3955 sequences, of which 807 are crossing events.

4.3. Training

For training, we use RMSProp [75] optimizer with initial learning rate of 10^{-4} for PIE and 5×10^{-5} for JAAD and batch size of 8. We trained the model for 300 epochs and reduced the learning rate by a factor of 0.2 based on the performance on the validation set. We empirically set α, β and γ values to 0.6, 1 and 1 respectively. To deal with class imbalance for action prediction, we applied class weights based on the ratio of positive and negative samples.

4.4. Metrics

The results are reported for the two primary tasks: trajectory and action prediction with 0.5s observation length.

Trajectory prediction. Following [5, 9, 31], we use two common metrics: Average Displacement Error $\mathrm{ADE} = \frac{\sum_{i=1}^{N} \sum_{j=t+1}^{\tau} ||y_i^j - \hat{y}_i^j||_2}{N \times \tau}$ and Final Displacement Error $\mathrm{FDE} = \frac{\sum_{i=1}^{N} ||y_i^{t+\tau} - \hat{y}_i^{t+\tau}||_2}{N}$. ADE and FDE metrics are measured based on the center coordinates of the bounding boxes $[x_c, y_c]$. In addition, to measure the accuracy of bounding box predictions, we report average and final RMSE of bounding box coordinates and denote them as ARB and FRB respectively. All metrics are reported in pixels for 1s prediction length.

Action prediction. As in [8, 42], we use common binary classification metrics, namely $\mathrm{accuracy}$, Area Under Curve (AUC), $\mathrm{F1}$ and $\mathrm{precision}$.

4.5. Models

Trajectory Prediction. Some past ego-centric trajectory prediction algorithms [5, 35] are compared to well-known methods such as [31, 76], which are designed and tested on surveillance sequences that are different as they provide bird's eye view of scenes and are recorded using fixed cameras. As a result, we select methods that are trained and tested in a similar ego-centric setting as the proposed algorithms. These methods are **Future Person Localization (FPL)** [32], **Bayesian LSTM (B-LSTM)** [35], **FOL** [33], and two variations of the method introduced in [10], **PIE**$_{traj}$ which only uses bounding boxes for prediction and **PIE**$_{full}$ which is the complete multimodal model. FPL model predicts center coordinates of the bounding boxes, therefore we only report its results on $\mathrm{ADE/FDE}$ metrics.
Action Prediction. For action prediction, we report the re-

Table 1. Performance of the proposed method on the PIE dataset. ↑ and ↓ mean higher or lower values are better respectively.

	Method	$ADE\downarrow$	$FDE\downarrow$	$ARB\downarrow$	$FRB\downarrow$	$Acc\uparrow$	$AUC\uparrow$	$F1\uparrow$	$Prec\uparrow$
	FOL [33]	73.87	164.53	78.16	143.69	-	-	-	-
	FPL [32]	56.66	132.23	-	-	-	-	-	-
	B-LSTM [35]	27.09	66.74	37.41	75.87	-	-	-	-
	PIE_{traj} [10]	21.82	53.63	27.16	55.39	-	-	-	-
	PIE_{full} [10]	19.50	45.27	24.40	49.09	-	-	-	-
	ATGC [11]	-	-	-	-	0.59	0.55	0.36	0.35
	I3D [71]	-	-	-	-	0.79	0.75	0.64	0.61
	MM-LSTM [27]	-	-	-	-	0.84	0.84	0.75	0.68
	SF-GRU [8]	-	-	-	-	0.86	0.83	0.75	0.73
	PCPA [42]	-	-	-	-	0.86	0.84	0.76	0.73
Ours	**BiPed**	**15.21**	**35.03**	**19.62**	**39.12**	**0.91**	**0.90**	**0.85**	**0.82**
Ours	BiPed+NEP	18.03	43.01	23.26	46.96	0.90	0.89	0.83	0.78
Ours	BiPed-NFE	18.44	45.07	24.81	50.64	0.90	0.90	0.84	0.80

sults on state-of-the-art pedestrian crossing prediction algorithms, namely **ATGC** [11], **MM-LSTM** [27], **SF-GRU** [8], and **PCPA** [42], for which we pad the context sequence for compatibility with evaluation criteria. Given the similarity between action prediction and recognition tasks, we also use state-of-the-art action recognition model, **I3D** [71].

Data processing. A subset of algorithms mentioned above use optical flow and pose information. We use FlowNet 2.0 [77] pretrained on [78] for optical flow maps and OpenPose [79] pretrained on [80] for poses. All these features are generated offline.

4.6. PIE Dataset

4.6.1 Multitask Prediction

We follow the same evaluation protocol as in [9] and report the results for single models. For the proposed method, **Bifold Pedestrian (BiPed)** prediction, we report on the final model as well as variations of it with no future ego-motion information (*NFE*), and with a noisy ego-motion planner (*NEP*) for which a recurrent decoder, an LSTM similar to other decoders, is used to predict future ego-vehicle motion.

As illustrated in Table 1, our method, BiPed, achieves state-of-the-art performance on all metrics. For trajectory prediction, our method significantly improves the results compared to prior state-of-the-art PIE_{full} by up to 22% on ADE and 20% on ARB metrics. Significant improvements are also achieved on action prediction. Compared to PCPA, our model achieves 5% and 6% improvements on Acc and AUC and even more improvements on F1 and Prec by 9%. These results indicate that the proposed method has more balanced performance compared to others. As expected, when relying on noisy motion predictions, the performance of our model declines. However, it still outperforms PIE_{full} by 8% and 4% on ADE and ARB and PCPA by 4% and 7% on Acc and F1 respectively. It should be noted that our method still achieves state-of-the-art performance on most metrics even without future ego-motion information.

Our method is also more stable compared to the past arts. For instance, standard deviation of the method over 20 runs with random initialization on ADE is 0.21 compared to 0.46 for PIE_{full} and on F1 is 0.006 whereas for PCPA is 0.01. Qualitative examples are illustrated in Figure 4.

Figure 4. Qualitative results of the proposed algorithm on PIE. Trajectories show one second in future and correspond to ground truth, BiPed (**ours**), PIE_{full}, and B-LSTM. For actions, we report the results on BiPed (**ours**), SF-GRU and PCPA. The results correspond to pedestrian crossing (C) and not-crossing (NC) actions. Here, ground truth is denoted as GT.

Table 2. The impact of different encoding and decoding schemes. ↑ and ↓ mean higher or lower values are better respectively.

Modules	$ADE\downarrow$	$FDE\downarrow$	$ARB\downarrow$	$FRB\downarrow$	$Acc\uparrow$	$AUC\uparrow$	$F1\uparrow$	$Prec\uparrow$
MIE+MIP	15.87	35.26	20.61	39.53	0.85	0.86	0.76	0.67
MIE+MJP	16.22	36.41	21.71	42.22	0.89	0.89	0.82	0.77
MJE+MIP	15.73	35.40	21.29	41.07	0.86	0.87	0.78	0.70
MJE+MJP	16.53	36.36	22.51	43.47	0.87	0.88	0.80	0.72
MIE+MJP+MIP	15.59	35.68	20.52	40.86	0.90	0.89	0.83	0.80
All	**15.21**	**35.03**	**19.62**	**39.12**	**0.91**	**0.90**	**0.85**	**0.82**

4.6.2 Ablation study

Independent and joint processing. We examine the impact of different proposed encoding and decoding modules, namely Multimodal Independent Encoding (MIE), Multimodal Joint Encoding (MJE), Multitask Independent Prediction (MIP), and Multitask Joint Prediction (MJP).

As shown in Table 2, MIP and MJP play complementary roles, i.e. MIP results in better trajectory predictions while MJP improves the results on action prediction. Even though trajectories and actions are correlated, the manner in which they are learned together is important. Multiple trajectories can correspond to the crossing action as long as pedestrian and vehicle paths intersect. However, inferring trajectories from the crossing action is more ambiguous because information, such as direction or speed of the pedestrian, is not directly implied by the action. Such ambiguity can increase uncertainty of predicting trajectories when a joint prediction module is used. That is why the best performance on all metrics is achieved when both MIP and MJP methods are combined. In case of encoding modules, MIE and MJE, although when used individually the performance on different tasks do not vary much, when combined they tend to complement each other and boost performance on all metrics as shown in the last two rows of the table.

Interaction encoding. We examine the contribution of Categorical Interaction Module (CIM) to the overall performance of our method. Here, we only report on trajectory metrics since the variation on action prediction results were

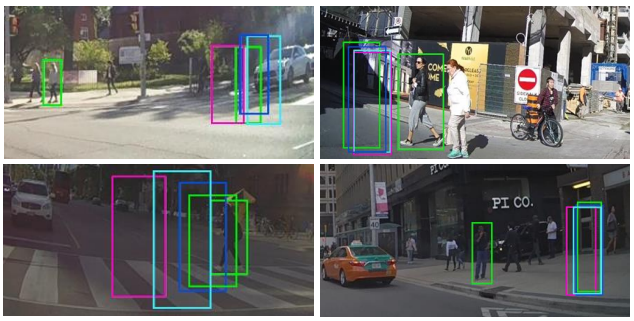

Figure 5. Qualitative results of the proposed algorithm on PIE showing the final predicted bounding boxes using no CIM, Hybrid and Hybrid+IAU CIM modules against ground truth.

Table 3. Evaluation of alternative interaction modeling methods. ↓ means lower values are better.

Interaction Model	$ADE\downarrow$	$FDE\downarrow$	$ARB\downarrow$	$FRB\downarrow$
No CIM	16.00	36.89	22.11	44.46
Single Conv2D	17.04	39.65	23.77	48.00
Single Conv3D	17.05	40.38	22.42	45.78
Categorical Conv2D	15.70	36.28	21.04	41.84
Categorical Conv3D	15.85	36.95	22.00	44.53
Ours (Single Hybrid)	16.17	37.17	21.39	42.21
Ours (Categorical Hybrid)	15.55	35.96	20.36	40.60
Ours (Categorical Hybrid+IAU)	**15.21**	**35.03**	**19.62**	**39.12**

Table 4. Ablation study on grid classification task. Grid cell (GC) sizes are in px^2. ↑ and ↓ mean higher or lower values are better respectively.

Method	$ADE\downarrow$	$FDE\downarrow$	$ARB\downarrow$	$FRB\downarrow$	$Acc\uparrow$	$AUC\uparrow$	$F1\uparrow$	$Prec\uparrow$
No grid	16.38	36.60	22.10	43.95	0.89	0.89	0.82	0.76
GC-15	16.30	35.88	21.41	42.82	0.90	**0.90**	0.84	0.78
GC-30	15.71	36.01	20.85	41.35	**0.91**	**0.90**	0.84	0.81
GC-60	**15.21**	**35.03**	**19.62**	**39.12**	**0.91**	**0.90**	**0.85**	**0.82**
GC-120	16.32	37.15	21.38	42.43	**0.91**	0.89	0.84	**0.82**

insignificant. We consider four alternative feature representation schemes where all classes are either represented in a *Single* semantic map or separated into *Categorical* maps and are processed using only 2D (*Conv2D*) or 3D (*Conv3D*) convolutional layers as specified in Figure 3. For Conv2D versions, we follow the method in [9] and average the maps along temporal dimension. We refer to our approach which uses both Conv2D layers and LSTMs as *Hybrid*.

The results are summarized in Table 3. Here, we can see that using CIM improves the results by up to 5% on ADE/FDE and 12% on ARB/FRB. The results also show the advantage of separating semantic maps into categories with shared characteristics. This can be due to the heterogeneous nature of pedestrians' interactions with their surroundings which necessitates learning a separate representation for each type of interaction. Overall, the proposed hybrid approach using categorical semantic representations clearly stands out on all metrics. Using the Interactive Attention Unit (IAU) further improves the results as the model learns to dynamically focus on different aspects of the interaction in a given context (see Figure 5 for some examples).

Grid classification task. We evaluate the proposed model with no auxiliary grid task (*no grid*) and different grid cell (*GC*) sizes in px^2. As shown in Table 4, overall, grid classification task is beneficial for both trajectory and action predictions, but improvements vary depending on the grid resolution. The best performance is achieved at grid cell size of 60 on all metrics, where ADE/ARB is improved by 7% and precision by 6%. When the grid cells are too large, e.g. 120, they are not effective, particularly on trajectories, which are affected the most because multiple time steps of the pedestrian movement may fall within a single cell.

Table 5. Performance of the proposed method on JAAD. ↑ and ↓ mean higher or lower values are better respectively.

	Method	$ADE\downarrow$	$FDE\downarrow$	$ARB\downarrow$	$FRB\downarrow$	$Acc\uparrow$	$AUC\uparrow$	$F1\uparrow$	$Prec\uparrow$
	FOL [33]	61.39	126.97	70.12	129.17	-	-	-	-
	FPL [32]	42.24	86.13	-	-	-	-	-	-
	B-LSTM [35]	28.36	70.22	39.14	79.66	-	-	-	-
	PIE_{traj} [10]	23.49	50.18	30.40	57.17	-	-	-	-
	PIE_{full} [10]	22.83	49.44	29.52	55.43	-	-	-	-
	ATGC [11]	-	-	-	-	0.64	0.60	0.53	0.50
	I3D [71]	-	-	-	-	0.82	0.75	0.55	0.49
	MM-LSTM [27]	-	-	-	-	0.80	0.60	0.40	0.39
	SF-GRU [8]	-	-	-	-	**0.83**	0.77	0.58	0.51
	PCPA [42]	-	-	-	-	**0.83**	0.77	0.57	0.50
Ours	**BiPed**	**20.58**	**46.85**	**27.98**	**55.07**	**0.83**	**0.79**	**0.60**	**0.52**
Ours	BiPed+NEP	20.75	47.44	28.16	55.50	0.83	0.79	0.60	0.51
Ours	BiPed+NFE	21.13	48.88	29.98	56.52	0.83	0.78	0.59	0.52

4.7. JAAD Dataset

We follow the same procedure as in Section 4.6.1 and evaluate our model on the JAAD dataset. As shown in Table 5, the performance improvement is smaller, particularly on action prediction, due to the fact that compared to PIE, the JAAD dataset is less diverse, less balanced, and does not contain the ego-vehicle motion information. However, for trajectory prediction, our model clearly stands out by improving over PIE_{full} by up to 10% on ADE and 5% on ARB, while maintaining state-of-the-art performance on action prediction with 2% improvement on AUC and F1 metrics. Using a noisy motion planner on JAAD, the performance decline is negligible because predictions are made on a small set of driver's actions as opposed to continuous velocity of the vehicle provided in PIE.

5. Conclusion

We presented a multitask learning framework for predicting pedestrian trajectory and action. Our method relies on a bifold mechanism to encode and decode different input modalities and tasks, thus allowing the model to learn cross-correlation between them and inducing it to learn better representations. In addition, we introduced a novel technique which implicitly models interactions between target pedestrians and their surroundings by relying on changes in semantic representations of the scenes. Using publicly available benchmarks, we showed that our proposed method significantly improves over existing methods on both trajectory and action prediction tasks. We further showed the overall contributions of our novel modules by conducting ablation studies. The proposed approach can be applied to human behavior understanding in other computer vision and robotics tasks, such as action and gesture recognition, interaction prediction, and group activity understanding.

References

[1] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Agreeing to cross: How drivers and pedestrians communicate," in *Intelligent Vehicles Symposium (IV)*, 2017.

[2] A. Rasouli and J. K. Tsotsos, "Autonomous vehicles that interact with pedestrians: A survey of theory and practice," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 900–918, 2019.

[3] J. Sun, Q. Jiang, and C. Lu, "Recursive social behavior graph for trajectory prediction," in *CVPR*, 2020.

[4] O. Makansi, O. Cicek, K. Buchicchio, and T. Brox, "Multimodal future localization and emergence prediction for objects in egocentric view with a reachability prior," in *CVPR*, 2020.

[5] S. Malla, B. Dariush, and C. Choi, "TITAN: Future forecast using action priors," in *CVPR*, 2020.

[6] M. Chaabane, A. Trabelsi, N. Blanchard, and R. Beveridge, "Looking ahead: Anticipating pedestrians crossing with future frames prediction," in *WACV*, 2020.

[7] B. Liu, E. Adeli, Z. Cao, K.-H. Lee, A. Shenoi, A. Gaidon, and J. C. Niebles, "Spatiotemporal relationship reasoning for pedestrian intent prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3485–3492, 2020.

[8] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Pedestrian action anticipation using contextual feature fusion in stacked RNNs," in *BMVC*, 2019.

[9] J. Liang, L. Jiang, J. C. Niebles, A. G. Hauptmann, and L. Fei-Fei, "Peeking into the future: Predicting future person activities and locations in videos," in *CVPR*, 2019.

[10] A. Rasouli, I. Kotseruba, T. Kunic, and J. K. Tsotsos, "PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction," in *ICCV*, 2019.

[11] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "Are they going to cross? A benchmark dataset and baseline for pedestrian crosswalk behavior," in *ICCVW*, 2017.

[12] M. Liu, S. Tang, Y. Li, and J. Rehg, "Forecasting human object interaction: Joint prediction of motor attention and actions in first person video," in *ECCV*, 2020.

[13] A. Piergiovanni, A. Angelova, A. Toshev, and M. S. Ryoo, "Adversarial generative grammars for human activity prediction," in *ECCV*, 2020.

[14] H. Joo, T. Simon, M. Cikara, and Y. Sheikh, "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction," in *CVPR*, 2019.

[15] T. Yao, M. Wang, B. Ni, H. Wei, and X. Yang, "Multiple granularity group interaction prediction," in *CVPR*, 2018.

[16] M. Strickland, G. Fainekos, and H. B. Amor, "Deep predictive models for collision risk assessment in autonomous driving," in *ICRA*, 2018.

[17] K.-H. Zeng, S.-H. Chou, F.-H. Chan, J. Carlos Niebles, and M. Sun, "Agent-centric risk assessment: Accident anticipation and risky region localization," in *CVPR*, 2017.

[18] D. Epstein, B. Chen, and C. Vondrick, "Oops! Predicting unintentional action in video," in *CVPR*, 2020.

[19] Y. Ma, X. Zhu, X. Cheng, R. Yang, J. Liu, and D. Manocha, "Autotrajectory: Label-free trajectory extraction and prediction from videos using dynamic points," in *ECCV*, 2020.

[20] M. Qi, J. Qin, Y. Wu, and Y. Yang, "Imitative non-autoregressive modeling for trajectory forecasting and imputation," in *CVPR*, 2020.

[21] P. Felsen, P. Agrawal, and J. Malik, "What will happen next? Forecasting player moves in sports videos," in *ICCV*, 2017.

[22] L. Fang, Q. Jiang, J. Shi, and B. Zhou, "TPNet: Trajectory proposal network for motion prediction," in *CVPR*, 2020.

[23] Y. Hu, S. Chen, Y. Zhang, and X. Gu, "Collaborative motion prediction via neural motion message passing," in *CVPR*, 2020.

[24] A. Mohamed, K. Qian, M. Elhoseiny, and C. Claudel, "Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction," in *CVPR*, 2020.

[25] H. Sun, Z. Zhao, and Z. He, "Reciprocal learning networks for human trajectory prediction," in *CVPR*, 2020.

[26] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *ECCV*, 2020.

[27] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "VIENA: A driving anticipation dataset," in *ACCV*, 2019.

[28] C. Choi and B. Dariush, "Looking to relations for future trajectory forecast," in *ICCV*, 2019.

[29] P. Zhang, W. Ouyang, P. Zhang, J. Xue, and N. Zheng, "SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction," in *CVPR*, 2019.

[30] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *CVPR*, 2019.

[31] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018.

[32] T. Yagi, K. Mangalam, R. Yonetani, and Y. Sato, "Future person localization in first-person videos," in *CVPR*, 2018.

[33] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *ICRA*, 2019.

[34] Y. Yao, M. Xu, Y. Wang, D. J. Crandall, and E. M. Atkins, "Unsupervised traffic accident detection in first-person videos," in *IROS*, 2019.

[35] A. Bhattacharyya, M. Fritz, and B. Schiele, "Long-term on-board prediction of people in traffic scenes under uncertainty," in *CVPR*, 2018.

[36] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *CVPR*, 2019.

[37] H. Zhao and R. P. Wildes, "Spatiotemporal feature residual propagation for action prediction," in *ICCV*, 2019.

[38] L. Chen, J. Lu, Z. Song, and J. Zhou, "Part-activated deep reinforcement learning for action prediction," in *ECCV*, 2018.

[39] Y. Shi, B. Fernando, and R. Hartley, "Action anticipation with RBF kernelized feature mapping RNN," in *ECCV*, 2018.

[40] Y. Kong, Z. Tao, and Y. Fu, "Deep sequential context networks for action prediction," in *CVPR*, 2017.

[41] M. Sadegh Aliakbarian, F. Sadat Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging LSTMs to anticipate actions very early," in *ICCV*, 2017.

[42] I. Kotseruba, A. Rasouli, and J. K. Tsotsos, "Benchmark for evaluating pedestrian action prediction," in *WACV*, 2021.

[43] K. Saleh, M. Hossny, and S. Nahavandi, "Real-time intent prediction of pedestrians for autonomous ground vehicles via spatio-temporal DenseNet," in *ICRA*, 2019.

[44] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *ICRA*, 2019.

[45] S. H. Park, G. Lee, M. Bhat, J. Seo, M. Kang, J. Francis, A. R. Jadhav, P. P. Liang, and L.-P. Morency, "Diverse and admissible trajectory forecasting through multimodal context understanding," in *ECCV*, 2020.

[46] J. Li, F. Yang, M. Tomizuka, and C. Choi, "EvolveGraph: Multi-agent trajectory prediction with dynamic relational reasoning," in *NeurIPS*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., 2020.

[47] V. Kosaraju, A. Sadeghian, R. Martin-Martin, I. Reid, H. Rezatofighi, and S. Savarese, "Social-BIGAT: Multimodal trajectory forecasting using Bicycle-GAN and graph attention networks," in *NeurIPS*, 2019.

[48] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *ICML*, 2018.

[49] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D pose estimation and action recognition using multitask deep learning," in *CVPR*, 2018.

[50] M. Guo, A. Haque, D.-A. Huang, S. Yeung, and L. Fei-Fei, "Dynamic task prioritization for multitask learning," in *ECCV*, 2018.

[51] G. Hu, L. Liu, Y. Yuan, Z. Yu, Y. Hua, Z. Zhang, F. Shen, L. Shao, T. Hospedales, N. Robertson, and Y. Yang, "Deep multi-task learning to recognise subtle facial expressions of mental states," in *ECCV*, 2018.

[52] K. Du, X. Lin, Y. Sun, and X. Ma, "CrossInfoNet: Multi-task information sharing based hand pose estimation," in *CVPR*, 2019.

[53] A. Mallya, D. Davis, and S. Lazebnik, "Piggyback: Adapting a single network to multiple tasks by learning to mask weights," in *ECCV*, 2018.

[54] Z. Zeng, X. Li, Y. K. Yu, and C.-W. Fu, "Deep floor plan recognition using a multi-task network with room-boundary-guided attention," in *ICCV*, 2019.

[55] Z. Tang, M. Naphade, S. Birchfield, J. Tremblay, W. Hodge, R. Kumar, S. Wang, and X. Yang, "PAMTRI: Pose-aware multi-task learning for vehicle re-identification using highly randomized synthetic data," in *ICCV*, 2019.

[56] K. Hassani and M. Haley, "Unsupervised multi-task feature learning on point clouds," in *ICCV*, 2019.

[57] S. Casas, W. Luo, and R. Urtasun, "IntentNet: Learning to predict intention from raw sensor data," in *CORL*, 2018.

[58] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *CVPR*, 2018.

[59] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3D object detection," in *CVPR*, 2019.

[60] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *CVPR*, 2019.

[61] P. Wu, S. Chen, and D. N. Metaxas, "MotionNet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps," in *CVPR*, 2020.

[62] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *CVPR*, 2018.

[63] X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu, "A modulation module for multi-task learning with applications in image retrieval," in *ECCV*, 2018.

[64] Y. Gao, J. Ma, M. Zhao, W. Liu, and A. L. Yuille, "NDDR-CNN: Layerwise feature fusing in multi-task cnns by neural discriminative dimensionality reduction," in *CVPR*, 2019.

[65] F. J. Bragman, R. Tanno, S. Ourselin, D. C. Alexander, and J. Cardoso, "Stochastic filter groups for multi-task cnns: Learning specialist and generalist convolution kernels," in *ICCV*, 2019.

[66] 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.

[67] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "GD-GAN: Generative adversarial networks for trajectory prediction and group detection in crowds," in *ACCV*, 2019.

[68] Z. Zhang, J. Gao, J. Mao, Y. Liu, D. Anguelov, and C. Li, "STINet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction," in *CVPR*, 2020.

[69] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv:1706.05098*, 2017.

[70] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv:1508.04025*, 2015.

[71] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *CVPR*, 2017.

[72] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv:1706.05587*, 2017.

[73] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The CityScapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[74] A. Rasouli, I. Kotseruba, and J. K. Tsotsos, "It's not all about size: On the role of data properties in pedestrian detection," in *ECCVW*, 2018.

[75] T. Tieleman and G. Hinton, "Lecture 6.5-RMSProp, coursera: Neural networks for machine learning," Tech. Rep., 2012.

[76] 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.

[77] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *CVPR*, 2017.

[78] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016.

[79] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *CVPR*, 2017.

[80] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.