Action-based Contrastive Learning for Trajectory Prediction

Marah Halawa, Olaf Hellwich, and Pia Bideau

Excellence Cluster Science of Intelligence, Technische Universitat Berlin
{marah.halawa,olaf.hellwich,p.bideau}@tu-berlin.de

Abstract. Trajectory prediction is an essential task for successful human-robot interaction, such as in autonomous driving. In this work, we address the problem of predicting future pedestrian trajectories in a first-person view setting with a moving camera. To that end, we propose a novel action-based contrastive learning loss, that utilizes pedestrian action information to improve the learned trajectory embeddings. The fundamental idea behind this new loss is that trajectories of pedestrians performing the same action should be closer to each other in the feature space than the trajectories of pedestrians with significantly different actions. In other words, we argue that behavioral information about pedestrian action influences their future trajectory. Furthermore, we introduce a novel sampling strategy for trajectories that is able to effectively increase negative and positive contrastive samples. Additional synthetic trajectory samples are generated using a trained Conditional Variational Autoencoder (CVAE), which is at the core of several models developed for trajectory prediction. Results show that our proposed contrastive framework employs contextual information about pedestrian behavior, *i.e.* action, effectively, and it learns a better trajectory representation. Thus, integrating the proposed contrastive framework within a trajectory prediction model improves its results and outperforms state-of-the-art methods on three trajectory prediction benchmarks.

1 Introduction

Predicting the future trajectories of pedestrians is an important task in many applications, such as in social robot interaction and autonomous driving. Typically, the future trajectory of an agent/pedestrian is predicted based on its own past movement history [33]. Nonetheless, integrating additional information is possible, such as the trajectories of surrounding agents [1, 9], or visual scene data [34]. When the surrounding agents in the scene are cars or robots, modeling the motion information based on past trajectories only is a reasonable way to solve the task. However, in this work, we argue that when other agents in the scene are pedestrians, then limiting the information used for prediction to past trajectories is not sufficient. In those cases additional information about pedestrian behavior (*e.g.* action) plays an important role for predicting their future trajectory. For example, the future trajectories of a pedestrian who is walking while texting on

a phone could be different from a pedestrian carrying an object or pushing a baby stroller even if they have the same previous observed trajectories, and the same end goal.

In this work, we study the influence of observed pedestrians' actions on their predicted trajectories. We propose a novel contrastive learning loss called *Action-based Contrastive Loss*. This novel loss is employed as a regularizer to the main trajectory prediction loss. The action-based contrastive loss encourages the trajectory embeddings of agents performing the same action (called positive samples) to come closer to each other in the feature space, and the embeddings of trajectories observed while performing different actions (called negative samples) far away from each other. For instance, the representations of trajectories of walking pedestrians are encouraged to become closer in the feature space, but farther from the representations of trajectories of pedestrians riding bikes or standing, as illustrated in Fig. 1.

Contrastive learning losses, including ours (action-based contrastive loss), utilize a mechanism called negative sampling/mining, which aims to choose the samples that are deemed different and therefore their corresponding features are driven farther in the embedding space. In our case, the negatives are trajectories of pedestrians that have different actions. Commonly used negative sampling techniques include choosing all other samples from the same mini-batch [5] or from a fixed-size memory bank [14]. Nevertheless, while these mechanisms prove effective on natural imaging datasets, we find they do not provide similarly high gains on trajectory datasets. We conjecture that this is due to the higher variation in visual data compared to trajectory data, and most importantly, to the larger sizes of imaging datasets, *e.g.* Imagenet [8] contains 1.6M images compared to PIE [31] that contains 738,970 trajectory samples. This results in limited numbers of negative samples, an issue that becomes more evident when conditioning samples by class information, *e.g.* action or behavior. Few works attempt to address this issue via designing special heuristics for negative mining [23, 38, 13]. Alternatively, in this work, we propose to utilize the data distribution learned by a Conditional Variational Auto-Encoder (CVAE) [35]. This avoids designing special heuristics for negative mining. While this form of sampling may be utilized to create negative samples only, we employ it to create both positive and negative samples. This is possible due to the different definition of our contrastive loss compared to the traditional Noise Contrastive Estimation loss (NCE loss); the notion of positive/negative in our case is tied to the different classes of action in the data. As explained above, the samples that belong to the same action class are positives from the point of view of this class, and other samples are negative.

Contributions. Our main contributions in this paper are as follows:

- A novel contrastive loss, called action-based contrastive loss, which provides the model with additional information about the action of an agent by guiding the development of the embedding space for trajectories during learning.

– A novel sampling/mining technique that utilizes the latent trajectory distributions learned by CVAEs, circumventing the need to design special mechanisms based on heuristics.

Our proposed contrastive learning framework improves the performance results on three first-person view trajectory prediction benchmarks. It also provides evidence that utilizing agent behavior information, in the form of action type in this case, is beneficial for trajectory prediction, aligning with [26]. However, our proposed learning framework requires action information only during training.

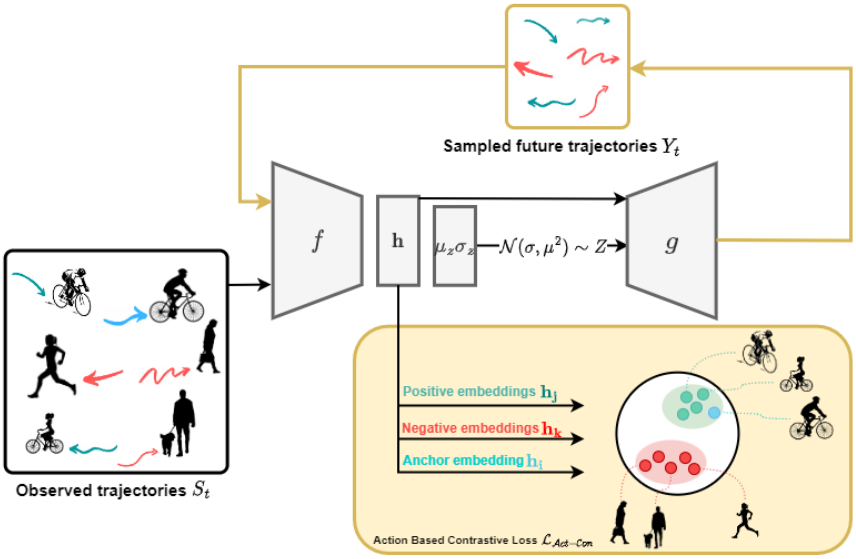

Fig. 1. Overview of our action-based contrastive learning framework during training phase. The contrastive loss $\mathcal{L}_{\text{Act-Con}}$ gets as input both positive (green) and negative (red) embeddings h for an anchor (blue). The positive and negative samples are the samples other than the anchor in the batch, as well as the synthetic samples from the CVAE. The parts shown in yellow refer to our novel action-based contrastive learning framework. It is worth mentioning that the action-based contrastive loss illustrated in this figure updates only the weights of encoder f, and it is jointly optimized with $\mathcal{L}_{traj}$ that updates both encoder f and decoder g. $\mathcal{L}_{traj}$ is not shown in the figure.

2 Related Work

Multi-modal trajectory prediction: A human can reach a desired location following many possible trajectories. Therefore, multiple works utilize multi-modal trajectory models, instead of predicting a single-path solution. Lee *et*

al. [21] proposed multi-modal trajectory model by incorporating samples from the Gaussian distribution of a trained conditional variational autoencoder (CVAE) into a long short-term memory encoder-decoder (LSTMED) model. Mangalam *et al.* [27] predict the multi-modal trajectory of an agent by modeling three factors: the desired endpoint goal, the social interaction with other agents in the scene, and the planned trajectories with respect to the environmental constraints in the scene. Similarly, their model is based on CVAE, which takes as input both the encodings of the past trajectory and of the endpoint goal. Sadeghian *et al.* [34] additionally include the past/observed trajectories of all agents for future trajectory prediction. To provide additional context information, top view images are incorporated. The distribution over feasible future paths is modeled for each agent using LSTM-based GAN module. Similarly, Yao *et al.* [39] predict trajectories conditioned on an estimated goal using a bi-directional RNN decoder. While our method has the potential of being added to any trajectory prediction method, we base our contrastive framework on BiTraP [39], in the first-person view setting.

Using human actions to improve trajectory prediction: In literature, many works employed video data to predict human activities [30]. Montes *et al.* [28] used a 3D-CNN as a feature extraction network then pass the learned representation to an RNN to exploit the time component in video data effectively. Ma *et al.* [24] improved the performance of LSTMs in human activity prediction by implementing ranking losses that penalize the prediction model on inconsistency in prediction scores from the sequence frames. Liang *et al.* [22] predicted a pedestrian's future trajectory simultaneously with future activities in a multi-task learning scheme. Rasouli *et al.* [31], studied the influence of an estimated pedestrian intention on the predicted trajectory by combining the intention representation with the observed trajectory coordinates, then used this representation as input to the decoder. Malla *et al.* [26] incorporates pedestrian action information with a trajectory prediction model. They require this information as prior information and learn a joint representation for both observed trajectory and pedestrian action. In this work, we also highlight the importance of analyzing the pedestrian's behavior and action in the prediction of their future trajectory. However, we propose to incorporate action information only during training using a novel action-based contrastive loss.

Contrastive trajectory prediction: Contrastive learning is a representation learning approach, first proposed by [29]. This approach encourages similar high-dimensional input vectors to be mapped closely to each other in a lower-dimensional embedding manifold, and the dissimilar ones are mapped far away from each other. Contrastive learning has been applied in several unsupervised [29, 15, 5, 11, 41, 4, 14, 6, 10, 17] and supervised [18] representation learning methods. Recently, only few works applied contrastive learning to trajectory prediction in a multi-agent setting. The flexibility of defining a contrastive loss by using positive and negative samples addresses the shortage problem in critical and challenging scenarios in training datasets. Such rare scenarios are necessary for the model, as the agent could face these in the real-world. Makansi *et al.* [25]

utilize this idea by separating the hard and critical samples in the feature space that do not satisfy some certain favorable criterion far away from the positive easy samples. Liu *et al.* [23] proposed a social sampling strategy that relies on augmenting negative samples with prior knowledge about undesired scenarios in the multi-agent setting. Both methods use the contrastive loss as a weighted combination to the future trajectory forecasting loss, which may be the mean squared error (MSE) or negative log-likelihood (NLL). Our method follows this family of algorithms, and uses a novel action-based contrastive loss to add context information about pedestrian actions to the trajectory prediction model.

Supervised contrastive loss: Khosla *et al.* [18] proposed a supervised contrastive loss that is a generalization of the Triplet loss [16]. In this supervised contrastive loss for each anchor there are more than one positive sample, in addition to many negative samples. There are two major differences between our proposed action-based contrastive loss and the supervised contrastive loss used in [18]. First, they employ the supervised contrastive loss to *replace* the cross-entropy loss for training the image classifier using image labels. However, we utilize the contrastive loss to *regularize* the trajectory prediction loss, which may be MSE or NLL. Second, due to the differences between the nature of datasets we use in this paper and the image data used in [18], it is simpler to extract many positive and negative samples from a large dataset, such as ImageNet [8]. However, in first-person view trajectory prediction datasets, the number of pedestrians with same actions is limited, therefore we address this with a novel sampling process from a CVAE trained to predict trajectories based on observing a short past trajectory. This CVAE predictive model ensures consistency between observed and predicted trajectories. Thus, it allows sampling additional positive and negative samples that belong to specific actions. Using this novel sampling technique avoids designing hard negative mining techniques, which use heuristics, as in [36, 19] for domain adaptation.

3 Methodology

In this section, we present our method for the task of pedestrian trajectory prediction, that focuses on integrating contextual information such as actions for more reliable future predictions. We address this by employing an action-based contrastive loss that enhances the trajectory prediction model with action information.

3.1 Problem Formulation

For each pedestrian we have an observed past trajectory $S_t = [s_1, ..., s_{t-1}, s_t]$ at time t, and we predict a future trajectory $Y_t = [y_{t+1}, y_{t+2}, ..., y_T]$, where s and y are bounding box coordinates for the observed and predicted trajectories, respectively. T is the maximum predicted trajectory time length in the future. In addition, we also have for each trajectory the action class information a, where the set of available actions $a \in \{a_1, a_2, ..\}$ may vary across different datasets.

Then in the training data, we assume there are N different training samples, where for each sample $i \in [1, .., N]$, we know S^i, Y^i, and a^i. Finally, we process the dataset samples in mini-batches, where each batch contains B samples.

3.2 Multi-modal Trajectory Prediction

We follow the commonly used approach of an encoder-decoder prediction model, where an encoder f learns the representation h given an observed trajectory S_t as an input, then a decoder g uses the representation h together with a sampled latent variable z to predict the future trajectory Y_t. We employ a standard long-short term encoder-decoder model (LSTMED) [21]. In fact, we extend on the bi-directional version of LSTMED, proposed in Yao *et al.*'s BiTraP [39]. The possibility to draw multiple future trajectories for each observed trajectory is achieved with a CVAE, which is a non-parametric model, that learns the distribution of target trajectory through a stochastic latent variable. The distribution learned by the CVAE is essential for our proposed contrastive framework, which we explain below. As a trajectory prediction loss function $\mathcal{L}_{traj}$, the Best-of-Many (BoM) L2-loss [3] between predicted and target trajectory is used. It is noteworthy that we do not restrict our proposed framework, explained below, to these choices of model architectures or loss functions; we adopt standard and effective techniques to study its influence on predicted trajectories. The essential factor for our learning framework is that the predicting future trajectory model is based on CVAE, similar to trajectory prediction models in [39, 27].

3.3 Action-based Contrastive Learning Framework

In order to enhance the model with contextual information about the pedestrian actions, we propose a novel loss that is called action-based contrastive loss, which acts as a regularizer for the trajectory prediction loss, and they jointly train the trajectory prediction model. The proposed action-based contrastive loss is based on a novel action-based sampling strategy shown in Fig. 1. We first describe the proposed contrastive loss in the simple case, without including additional samples from the CVAE distribution, and we generalize it later.

Action-based Contrastive Loss Let B be the number of samples within a batch. For each observed past trajectory S^i where $i \in \{1, .., B\}$, called the anchor, there exists multiple positive and negative samples. The positive samples S^{i+} are the trajectories that have the same action class as the anchor, which are denoted by $S^{i''}$. Moreover, we also add an augmented version $S^{i'}$ of the anchor trajectory as a positive sample, following [23], which is created by adding small white noise ϵ to the bounding box coordinates of the anchor trajectory.

Formally:

$$S^{i'} = \left\{S^i + \epsilon\right\}$$
$$S^{i''} = \left\{S^j\right\}; \text{where } 0 < j < B, a^j = a^i, i \neq j$$
$$S^{i+} = S^{i'} \cup S^{i''}$$

Negative samples S^{i-} are trajectories belonging to a different action class than the anchor.

$$S^{i-} = \left\{S^k\right\}; \text{where } 0 < k < B, a^k \neq a^i, i \neq k$$

Afterwards, all batch samples $\left\{S^i\right\}_{i=1}^{B}$ are processed by the model encoder f to produce their hidden representations $\left\{\mathbf{h}^i\right\}_{i=1}^{B}$. Assuming M positive samples and K negative samples in the batch, with $B = M + K$. The proposed loss is calculated as follows:

$$\begin{aligned} \ell_{\text{Act-Con}} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\sum_{j=1,j\neq i,a^j=a^i}^{M}\exp(sim(\mathbf{h}^i,\mathbf{h}^j)/\tau)}{\sum_{k=1,k\neq i}^{K}\exp(sim(\mathbf{h}^i,\mathbf{h}^k)/\tau)} \\ \mathcal{L}_{\text{Act-Con}} &= \frac{1}{N/B}\sum^{N/B}\ell_{\text{Act-Con}} \end{aligned} \tag{1}$$

where sim is the similarity between the vector representations of the samples, for which we use the dot-product. τ is the temperature hyperparameter. The above loss function encourages the embeddings $\mathbf{h}^i$ of positive sample trajectories to be closer to each other in the embedding space, and far away from the embeddings of negative samples. The complete loss function sums both the trajectory prediction loss $\mathcal{L}_{traj}$ and the action-based contrastive loss $\mathcal{L}_{\text{Act-Con}}$:

$$\mathcal{L}_{final} = \mathcal{L}_{traj} + \beta\mathcal{L}_{\text{Act-Con}} \tag{2}$$

where β is a hyper-parameter that controls the contribution of action-based contrastive loss. It is worth mentioning that additional behavioral information such as pedestrian's action class is only needed during training. However, during inference, the model only takes the observed trajectory as input to predict the future trajectory.

Action-based Synthetic Trajectory Sampling The above loss formulation assumes no additional synthetic samples, *i.e.* it considers observed trajectories in the batch only. However, due to the relatively limited sizes of trajectory datasets, and the shortage of diversity in action classes in captured scenes, commonly used negative sampling techniques may not be sufficient. Those include sampling from the same mini-batch [5] or from a fixed-size memory bank [14]. More comprehensive negative and positive samples, from various behavior scenarios are rather needed. Therefore, we extend training samples by drawing trajectories from the distribution learned by the generative Conditional Variational Autoencoder (CVAE) model. CVAE is a generative model that introduces a stochastic latent variable Z in order to learn the distribution of target future trajectory $P\left(Y^i|S^i,Z\right)$. This distribution is conditioned on the input observed trajectories S^i, and the stochastic latent variable Z. Thus, the model is able to predict *multiple* feasible trajectories Y^i given the input S^i. We assume the latent variable following a Gaussian distribution $Z \sim N\left(\mu_Z, \sigma_Z^2\right)$, and we train the CVAE

to capture this distribution. Afterwards, the training dataset is extended by sampling from the Gaussian latent space multiple times, and passing samples through the decoder g to effectively predict different feasible future trajectories conditioned on an observed trajectory. The conditioning on the observed trajectory ensures a consistent behavior in the predicted future trajectory. This behavior is captured in both the continuity of the trajectory as well as the identical action class in both observed and future trajectories. We employ the same encoder-decoder trajectory prediction model explained above in Sec. 3.2, which is a CVAE that predicts multiple feasible future trajectories, as the example in Fig. 2 shows. This sampling strategy is illustrated in Fig. 1, and it has the advantage that it avoids designing heuristics for negative sample mining techniques, as mentioned before. The intuition behind this sampling strategy is that the encoder-decoder CVAE model is capable of generating future trajectories with the same behavior of the observed trajectory. Since it is trained to predict the future trajectory of an observed trajectory, then it captures the characteristics of the observed trajectory.

Let $\left\{Y^{i,l}\right\}_{l=1}^{L}$ be the multiple predicted trajectories for an observed trajectory S^i. Here, $Y^{i,l}$ is sampled from $P\left(Y^{i,l}|S^i, Z\right)$, and L is the number of times we sample a different Z from the normal distribution $N\left(\mu_Z, \sigma_Z^2\right)$. Given these synthetic trajectory samples, the set of positive samples for trajectory S^i are then reformulated as follows:

$$S_{1:t}^{i+} = \left\{S^{i'}{}_{1:t}\right\} \cup \left\{S^{i''}{}_{1:t}\right\} \cup \left\{Y_{t+1:T}^{j,l}\right\}_{l=0}^{L}$$

where $i, j \in 1, .., B$ and $a^l = a^i$ and $a^j = a^i$. And the negative samples are reformulated as follows:

$$S_{1:t}^{i-} = \left\{S_{1:t}^{k}\right\} \cup \left\{Y_{t+1:T}^{k,l}\right\}_{l=0}^{L}$$

where $i, k \in 1, .., B$ and $i \neq k$ and $a^i \neq a^k$ and $a^i \neq a^l$ and $a^k = a^l$. In words, the synthetic samples created for sample S^k, which we denote $Y^{k,l}$, are considered negative from the point of view of sample S^i. These synthetic samples have the same action class of sample S^k, hence denoted $a^k = a^l$.

The described action-based synthetic trajectory sampling strategy changes the sets of positive and negative samples used in creating training batches. However, the proposed contrastive loss equation Eq. 1 remains the same, only M and K are affected.

4 Experiments

In this section, we present the evaluation results of our method on three first-person-view trajectory prediction datasets [31, 32, 26]. First, we describe the used datasets. Then, we provide an overview of the experimental setup and used evaluation metrics. Finally, we discuss our results and findings.

Fig. 2. Examples from TITAN dataset showing the multi-modality in the trajectory prediction space. CVAE is able to predict multiple feasible future trajectories (Green bounding boxes), conditioned on previously observed trajectories (Blue bounding boxes). The red bounding boxes refer to the ground truth future trajectories.

4.1 Datasets

We evaluate our method on first-person view datasets. In this domain, the Pedestrian Intention Estimation (PIE) [31] and the Joint Attention for Autonomous Driving (JAAD) [32] datasets are the most commonly used benchmarks in literature. The PIE dataset provides 293,437 annotated frames, containing 1,842 pedestrians with behavior annotations such as walking, standing, crossing, looking, etc. Since a pedestrian could be "walking" and "looking" at the same time, for example, then a pedestrian could have multiple behavior labels in a single frame. Therefore, we only use two classes "walking" and "standing", which are exclusive. We use the same train and test splits in [31]. On the other hand, the JAAD dataset provides 82,032 annotated frames, containing 2,786 pedestrians, 686 of them have behavior annotations. Similar to the PIE dataset, we only use for JAAD dataset two classes "walking" and "standing", which are exclusive. We use the same train and test splits in [32].

We also use a third dataset named TITAN [26], which contains more action classes compared to PIE and JAAD. TITAN provides 75,262 frames with 395,770 pedestrians with multiple action labels organized in five hierarchical contextual activities, such as individual atomic actions, simple scene contextual actions, complex contextual actions, transportive actions, and communicative actions. For the same reason of not having multiple labels for each pedestrian, we use individual atomic actions labels for the TITAN dataset. The atomic action labels describe the primitive action, and are categorized into 9 labels (sitting, standing, walking, running, bending, kneeling, squatting, jumping, laying down).

4.2 Experimental Setup

We use the same setup for all datasets, where we observe 0.5 seconds and predict 0.5, 1.0, and 1.5 seconds, following [31, 39]. The predicted trajectories have two forms: bounding boxes coordinates and centers, that are evaluated separately.

Table 1. The quantitative results on **PIE** and **JAAD** datasets. The evaluation metrics are reported for different prediction lengths in *squared pixels*. ABC+ is our proposed action-based contrastive framework with sampling from a learned CVAE. BiTraP is a baseline trajectory prediction model without adding any contrastive loss. The other baseline results are obtained from [39]. Lower is better.

Method	**PIE**					**JAAD**				
	ADE			C-ADE	C-FDE	ADE			C-ADE	C-FDE
	0.5	1.0	1.5	1.5		0.5	1.0	1.5	1.5	
Linear [31]	123	477	1365	950	3983	233	857	2303	1565	6111
LSTM [31]	172	330	911	837	3352	289	569	1558	1473	5766
B-LSTM [2]	101	296	855	811	3259	159	539	1535	1447	5615
FOL-X [40]	47	183	584	546	2303	147	484	1374	1290	4924
PIE_{traj} [31]	58	200	636	596	2477	110	399	1280	1183	4780
BiTraP [39]	23	48	102	81	261	**38**	94	222	177	565
ABC+	**16**	**38**	**87**	**65**	**191**	40	**89**	**189**	**145**	**409**

PIE and JAAD datasets are both annotated at 30Hz frequency, therefore we observe 15 frames and predict 45 frames. However, TITAN dataset is annotated at 10HZ sampling frequency. Thus, we observe 5 frames and predict 15 frames.

Implementation details. We use 256 as the size for all hidden layers in the encoder-decoder model that is detailed in Sec. 3.2. It is noteworthy that we implement the loss in Eq. 1 using the efficient matrix-form (especially on GPU machines), instead of performing expensive pairwise computations. We train the model on all datasets with Adam optimizer [20] using a batch size of 128 and a learning rate of 0.001. On training datasets, we perform hyper-parameter tuning for β (Eq. 2). We achieve our best results using $\beta = 0.75$ for all datasets.

Evaluation metrics. Following the commonly used evaluation protocols in literature [31, 39, 37], we use the following evaluation metrics: i) Bounding box Average Displacement Error (ADE), ii) Bounding box Center ADE (C-ADE), iii) Bounding box Final Displacement Error (FDE), and iv) Bounding box Center FDE (C-FDE). All are computed in squared pixels. The bounding box ADE is the mean square error (MSE) for all predicted trajectories and ground-truth future trajectories. This error is calculated using the bounding box upper-left and lower-right coordinates. However, in C-ADE, otherwise called C-MSE, the error is calculated using the centers of the bounding boxes. Bounding box FDE, otherwise called FMSE, is the distance between the destination point of the predicted trajectory and of the ground truth at the last time step. FDE is also calculated using the bounding boxes coordinates. Finally, C-FDE or C-FMSE is the mean squared error between the centers of final destination bounding boxes.

Table 2. The quantitative results on **TITAN** dataset. The evaluation metrics are reported for observing 15 time steps and predicting 45 time steps of trajectories in *squared pixels*. ABC+ is our proposed action-based contrastive framework with sampling from a learned CVAE. BiTraP is a baseline trajectory prediction model without adding any contrastive loss. Lower is better.

Method	ADE			C-ADE	C-FDE
	0.5	1.0	1.5	1.5	
BiTraP [39]	194	352	658	498	989
ABC+	**165**	**302**	**575**	**434**	**843**

Table 3. The quantitative results on **TITAN** dataset. The evaluation metrics are reported for observing 10 time steps and predicting 20 time steps of trajectories in *pixels*. ABC+ is our proposed action-based contrastive framework with sampling from a learned CVAE. The other baseline results are obtained from [26]. Lower is better.

Method	ADE	FDE
Social-LSTM [1]	37.01	66.78
Social-GAN [12]	35.41	69.41
Titan-vanilla [26]	38.56	72.42
Titan-AP [26]	33.54	55.80
ABC+	**30.52**	**46.84**

Baselines. The trajectory prediction model trained with our proposed Action-Based Contrastive framework (loss and sampling strategy) is indicated by (ABC+). First, we evaluate the performance of our action-based contrastive framework by comparing its results to the original BiTraP trajectory prediction model [39] on all datasets, *i.e.* without adding our contrastive loss. This baseline aims to highlight the gains obtained by our proposed contrastive framework. BiTraP had previously achieved state-of-the-art on PIE and JAAD datasets. Additionally, for PIE and JAAD datasets, we compare our results with PIEtraj [31], FOL-X [40], B-LSTM [2], LSTM [31], and Linear [31] trajectory prediction models. On the TITAN dataset, we first report the evaluation results compared to BiTraP using observed and predicted lengths equal to PIE and JAAD. However, to fairly compare our results on the TITAN dataset with prior work of Malla *et al.* [26], Social-LSTM [1], and Social-GAN [12], we follow the same experimental setup used in [26]. To that end, we retrain both the BiTraP baseline model, and our proposed model (ABC+) to predict 20 frames, after observing 10 frames, and we report our results using ADE and FDE in pixels, not in squared pixels.

4.3 Trajectory Prediction Results

The evaluation results are shown in Tab. 1 for PIE and JAAD. For TITAN, in Tab. 2 we show the evaluation results when observing 10 frames and predicting

20 frames, similar to [26], and in Tab. 3 we compare to BiTraP when observing 5 frames and predicting 10 frames. As the tables show, our method (ABC+) achieves superior performance compared to the baseline BiTraP, which does not use our action-based contrastive loss. This result highlights the effectiveness of adding the proposed contrastive objective and sampling strategy. Our proposed method also outperforms other baseline methods, with significant margins.

These evaluation results confirm the gains obtained by using our proposed action-based contrastive loss and sampling strategy. Utilizing action information with our contrastive approach exhibits improved performance across all evaluated benchmarks. In the TITAN dataset, particularly, the performance benefits appear larger. We believe this is due to TITAN's more comprehensive action class structure, compared to PIE or JAAD. In other words, a more diverse set of pedestrian action classes improves the learned embedding space by our action-based contrastive loss. Nevertheless, our method improves the obtained results even with a simpler binary action class structure in PIE and JAAD.

Another significant result is that our method (ABC+) outperforms the baseline Titan-AP [26] on TITAN, which incorporates the same action class information with observed trajectory information, and produces a combined embedding to predict the future trajectory. This indicates that our approach of supporting the trajectory prediction model with behavioral context information by using action-based contrastive loss is more effective than encoding the action classes in the embedding space representation.

Table 4. Ablation results on **PIE**, **JAAD**, and **TITAN** datasets. ABC+ is our proposed action-based contrastive framework with sampling from a learned CVAE. ABC uses our action-based contrastive loss but without sampling from a learned CVAE. SimCLR uses a normal batch contrastive loss instead of our proposed action-based contrastive loss. Lower is better.

	Method	ADE			C-ADE	C-FDE
		0.5	1.0	1.5	1.5	
PIE	SimCLR	26	67	163	125	399
	ABC	16	40	93	69	213
	ABC+	**16**	**38**	**87**	**65**	**191**
JAAD	SimCLR	50	124	273	211	608
	ABC	41	93	201	150	425
	ABC+	**40**	**89**	**189**	**145**	**409**
TITAN	SimCLR	255	506	999	773	1805
	ABC	188	345	634	488	951
	ABC+	**165**	**302**	**575**	**434**	**843**

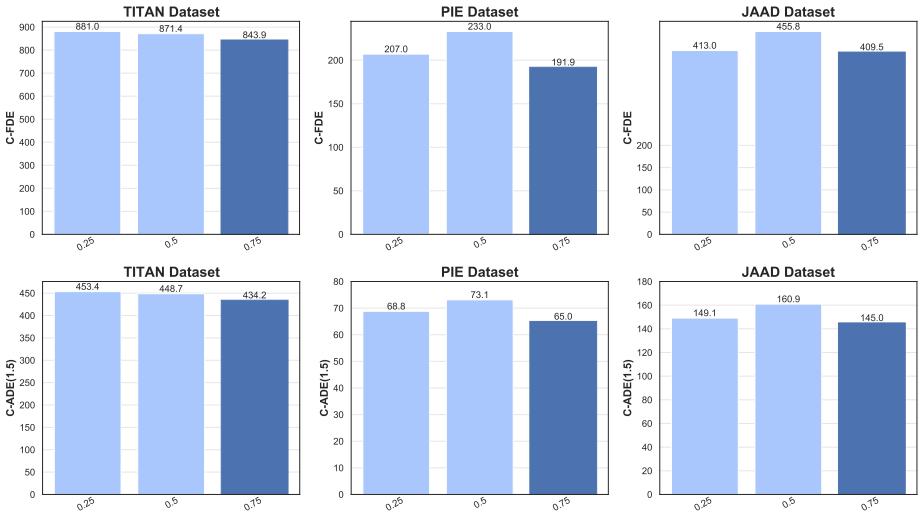

Fig. 3. C-FDE results (top) and C-ADE(1.5) (bottom) of trajectory prediction model by applying different β values 0.25,0.5,0.75 in Eq. 2. The results are reported for TITAN, PIE, and JAAD datasets. Lower values are better.

4.4 Ablation study

In this section, we present the ablation studies to provide further insights into our proposed action-based contrastive loss. Similar to the evaluation results shown above, we refer to our proposed action-based contrastive framework by ABC+, where we use the action-based contrastive loss as a regularizer to the trajectory prediction loss, and we also increase negative and positive samples during training by sampling synthetic trajectories from CVAE.

Does action information improve the contrastive loss? The first ablation study examines how the action-based contrastive loss Eq. 1 compares to the batch contrastive loss, namely SimCLR [5], shown in Tab. 4. For this baseline, we replace the action-based contrastive loss with SimCLR contrastive loss, and we measure the trajectory prediction performance. The results demonstrate the impact of utilizing contextual information in form of action on the future trajectory prediction model.

Is the proposed action-based sampling strategy using a CVAE effective? The second ablation study analyzes the impact of sampling synthetic trajectories from CVAE on the trajectory prediction model performance. The baseline ABC in Tab. 4 indicates the trajectory prediction model trained with action-based contrastive loss without using the extra synthetic samples from the learned CVAE. Comparing the quantitative results of ABC+ to the results of ABC highlights the effectiveness of our novel sampling strategy on all datasets.

How does the weight of the contrastive loss affect the results? Finally, we also study the influence of the hyper-parameter β in Eq. 2, which

controls the impact of the action-based contrastive loss into the final objective function. As shown in Fig. 3, we obtain the best results on all benchmark datasets by setting β to 0.75.

5 Discussion and Conclusions

We presented a contrastive framework for learning behavior-aware pedestrian trajectory representations. Our proposed framework consists of an action-based contrastive loss, and a novel trajectory sampling technique from a learned distribution of a C-VAE model. The proposed framework significantly improves the performance of trajectory prediction models on three different first-person view benchmarks. Our evaluation results provide evidence that including pedestrian behavior information, in the form of action or activity class in this case, is beneficial for trajectory prediction. Moreover, our results also confirm that our action-based contrastive loss, in conjunction with our sampling strategy, is superior to alternative approaches that also utilize action class information.

This work comes with a number of strengths. First, we ensure our proposed contrastive loss can be easily integrated with commonly used trajectory prediction models. Second, our proposed sampling strategy utilizes readily learned distributions by generative models, such as CVAEs, and it avoids designing data-specific heuristics. This allows for wider range of applications, such as on animal trajectory data. Finally, contrastive learning in general, and our proposed action-based framework, in particular, allow for enhancing the quantities of underrepresented action classes in the data. This line of work may help address the shortage of necessary edge-cases in training datasets, which may be encountered in real-world scenarios.

This work also comes with a limitation. While effective, our proposed action-based contrastive framework requires pedestrian action labels during the training phase only. However, this requirement is mitigated by our new trajectory sampling technique from CVAE, which does not require action labels for generated samples. Making the training scheme semi-supervised in our model. In addition, many modern trajectory datasets are increasingly providing action information. Action prediction tasks from video data achieve high performances [7], and hence can be performed efficiently and reliably as a pre-processing step for our trajectory prediction framework. We deem evaluating such idea as future work.

Acknowledgment

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–971 (2016). https://doi.org/10.1109/CVPR.2016.110
2. Bhattacharyya, A., Fritz, M., Schiele, B.: Long-term on-board prediction of people in traffic scenes under uncertainty. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4194–4202 (2018)
3. Bhattacharyya, A., Schiele, B., Fritz, M.: Accurate and diverse sampling of sequences based on a "best of many" sample objective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8485–8493 (2018)
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
6. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15750–15758 (2021)
7. Degardin, B., Proença, H.: Human behavior analysis: A survey on action recognition. Applied Sciences **11**(18) (2021). https://doi.org/10.3390/app11188324, https://www.mdpi.com/2076-3417/11/18/8324
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
9. Deo, N., Trivedi, M.M.: Convolutional social pooling for vehicle trajectory prediction. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1549–15498 (2018). https://doi.org/10.1109/CVPRW.2018.00196
10. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In: International Conference on Computer Vision (ICCV). pp. 9588–9597 (October 2021)
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning (2020)
12. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 2255–2264 (2018)
13. Harwood, B., Kumar B G, V., Carneiro, G., Reid, I., Drummond, T.: Smart mining for deep metric learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
15. Henaff, O.: Data-efficient image recognition with contrastive predictive coding. In: International Conference on Machine Learning. pp. 4182–4192. PMLR (2020)

16. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: International workshop on similarity-based pattern recognition. pp. 84–92. Springer (2015)
17. Jahanian, A., Puig, X., Tian, Y., Isola, P.: Generative models as a data source for multiview representation learning. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=qhAeZjs7dCL
18. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673. Curran Associates, Inc. (2020)
19. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: Self-supervised contrastive regularization for domain generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9619–9628 (2021)
20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 336–345 (2017)
22. Liang, J., Jiang, L., Niebles, J.C., Hauptmann, A.G., Fei-Fei, L.: Peeking into the future: Predicting future person activities and locations in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5725–5734 (2019)
23. Liu, Y., Yan, Q., Alahi, A.: Social nce: Contrastive learning of socially-aware motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15118–15129 (2021)
24. Ma, S., Sigal, L., Sclaroff, S.: Learning activity progression in lstms for activity detection and early detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1942–1950 (2016)
25. Makansi, O., Çiçek, O., Marrakchi, Y., Brox, T.: On exposing the challenging long tail in future prediction of traffic actors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13147–13157 (October 2021)
26. Malla, S., Dariush, B., Choi, C.: Titan: Future forecast using action priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11186–11196 (2020)
27. Mangalam, K., Girase, H., Agarwal, S., Lee, K., Adeli, E., Malik, J., Gaidon, A.: It is not the journey but the destination: Endpoint conditioned trajectory prediction. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12347, pp. 759–776. Springer (2020). https://doi.org/10.1007/978-3-030-58536-5_45, https://doi.org/10.1007/978-3-030-58536-5_45
28. Montes, A., Salvador, A., Pascual, S., Giro-i Nieto, X.: Temporal activity detection in untrimmed videos with recurrent neural networks. In: 1st NIPS Workshop on Large Scale Computer Vision Systems (December 2016)
29. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018)
30. Pareek, P., Thakkar, A.: A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artificial Intelligence Review **54**(3), 2259–2322 (2021)

31. Rasouli, A., Kotseruba, I., Kunic, T., Tsotsos, J.K.: Pie: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In: International Conference on Computer Vision (ICCV) (2019)
32. Rasouli, A., Kotseruba, I., Tsotsos, J.K.: Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 206–213 (2017)
33. Rudenko, A., Palmieri, L., Herman, M., Kitani, K.M., Gavrila, D.M., Arras, K.O.: Human motion trajectory prediction: a survey. The International Journal of Robotics Research **39**(8), 895–935 (2020). https://doi.org/10.1177/0278364920917446, https://doi.org/10.1177/0278364920917446
34. Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., Savarese, S.: Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
35. Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)
36. Sun, Y., Tzeng, E., Darrell, T., Efros, A.A.: Unsupervised domain adaptation through self-supervision. CoRR **abs/1909.11825** (2019), http://arxiv.org/abs/1909.11825
37. Wang, C., Wang, Y., Xu, M., Crandall, D.J.: Stepwise goal-driven networks for trajectory prediction. IEEE Robotics and Automation Letters **7**(2), 2716–2723 (2022). https://doi.org/10.1109/LRA.2022.3145090
38. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
39. Yao, Y., Atkins, E., Johnson-Roberson, M., Vasudevan, R., Du, X.: Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. IEEE Robotics and Automation Letters (RA-L) (2021)
40. Yao, Y., Xu, M., Choi, C., Crandall, D.J., Atkins, E.M., Dariush, B.: Egocentric vision-based future vehicle localization for intelligent driving assistance systems. 2019 International Conference on Robotics and Automation (ICRA) pp. 9711–9717 (2019)
41. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)