Trajectory Prediction with Contrastive Pre-training and Social Rank Fine-tuning

Chenyou Fan[1][0000−0002−9835−8507], Haiqi Jiang[1], Aimin Huang[0000−0001−9895−3202], and Junjie Hu[2][0000−0002−1911−4361]

[1] South China Normal University, Guangdong, China
fanchenyou@scnu.edu.cn
[2] Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRs), Shenzhen, Guangdong, China
hujunjie@cuhk.edu.cn

Abstract. This paper focuses on the accurate prediction of pedestrian trajectories in scenarios where individuals walk alone or in social groups, and sometimes alter their paths to avoid collisions. While previous work has improved backbone neural networks to model individual motion patterns, few studies have explicitly addressed the consistency of internal motion patterns or properness of external interactions. To address this, we propose a unified framework consisting of a Contrastive History-Prediction (CHIP) module and a Differentiable Social Interaction Ranking (DSIR) module. The CHIP module utilizes unsupervised contrastive loss to optimize predicted motion patterns consistent with observations, while the supervised DSIR module ensures predicted interactions are compatible with realistic positions. Our analysis and numerical studies demonstrate the effectiveness of our approach, which achieves a 5-10% improvement in positional accuracy and a 3-7% boost in interactive properness. We provide comprehensive visualizations of anticipated trajectories with temporal interactive scores across various scenarios.

Keywords: Trajectory Prediction · Contrastive Learning · Social Interaction.

1 Introduction

Predicting the future trajectories of autonomous vehicles is a critical task for safe navigation in dense urban traffic. Recent studies have made substantial progress in developing advanced deep-learning (DL) models to explore human movement patterns, such as using LSTMs [1], GANs [9, 20], Transformers [30], and GCNs [17, 23]. To model human interactions, these studies have proposed to aggregate neighbors' features with pooling [1, 15, 31], weighted averaging [17, 23], or multi-head attention [30].

In this study, instead of proposing new DL architectures, we focus on two fundamental aspects of trajectory prediction: the consistency of human movement patterns and the properness of human interactions. To this end, we propose a unified, model-agnostic training procedure to explicitly optimize motion consistency and quantify interaction properness. This approach allows us to better explain the interactive mechanisms implied by existing sophisticated DL models and to quantify the properness of the predicted social interactions.

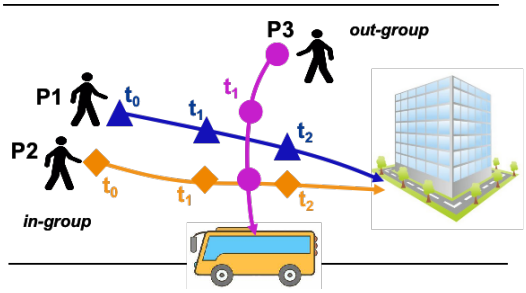


Fig. 1: Demo of pedestrian trajectory prediction. In a same social group, P1 and P2 keep close along the path. P3 is out-group w.r.t. P1 and P2, approaching at some future steps.

Human behaviors have been observed to exhibit stability and predictability, as people often follow consistent patterns in their movements. To investigate this phenomenon, this study focuses on examining the *internal consistency* of human behaviors. Our approach involves optimizing the similarity between observed motion patterns and predicted future patterns for each individual. This is achieved through contrastive learning, which associates a person's historical pattern with their corresponding future pattern. The proposed learning method, named *Contrastive HIstory-Prediction* (CHIP), aims to unsupervisedly ensure that predicted trajectories align with previously observed patterns. To accomplish this, motion embeddings are extracted from historical and future time steps for all individuals in the scene.

Humans are innately social beings who adjust their actions to facilitate appropriate social interactions. Our study identifies two significant types of pedestrian interactions, as in Fig. 1. The first type is called *in-group*, where individuals from the same social group tend to walk together and maintain a close proximity. The second type is *out-group*, where individuals walk separately but anticipate potential collisions in the near future. Consequently, they modify their trajectories using complex dynamics to maintain a comfortable social distance. Thus, the second crucial aspect of this research is to explicitly model the *external social properness* that dictate human behaviors.

In order to capture the *external social properness* of pedestrian behavior, we utilize a pairwise potential energy calculation based on the relative distance between individuals over time. We then create a spatial-temporal ranking of these potentials for all person pairs, reflecting the varying intensities of interactions. This ranking is demonstrated in Fig. 1, where at step t_1, the ranking of $(P1, P3)$ is higher than that of $(P2, P3)$ as they approach, with the order reversing at t_2. During training, the actual ranking is obtained from the ground-truth trajectories. The discrepancy between the predicted and actual ranking is used to determine the properness of predicted interactions, informing the design of a ranking loss which is optimized for the model end-to-end. This process is called Differential Social Interaction Ranking (DSIR), which aims to accurately capture the progression of social potentials in a supervised manner.

In summary, we propose explicitly modeling both *internal* and *external* factors of human behaviors as multi-task learning objectives. Notably, our CHIP and DSIR mod-

ules are both model-agnostic and parameter-free. We will demonstrate they can integrate into existing backbones seamlessly and improve prediction accuracy steadily.

In summary, the main contributions of our work include:

- We propose to learn internal movement and external interactive patterns to depict human behaviors in dense traffic;
- We apply unsupervised contrastive learning process to associate observed movements with predicted trajectories for pre-training;
- We design a ranking scheme to describe the dynamic interactions with pairwise potentials based on pedestrians' trajectories, and integrate into model optimization in an end-to-end manner;
- Our approach significantly outperforms existing methods by 5-10% in positional accuracy and 3-7% in interactive properness.

2 Related Work

Trajectory Prediction. In Autonomous Driving (AD) technical stack, trajectory prediction is an important perception task which aims to track mobile agents such as pedestrians and vehicles. Recent approaches commonly use RNNs [1, 9, 13, 27], GANs [9, 20] or GNNs [17, 23] to encode the history and decode to future trajectories. Recent AD studies [5, 7, 32] also utilize additional high-definition maps to refine the generation of future coordinates.

Social Interaction. Social-LSTM [1] modeled the interactions by pooling neighboring agents' features. Social-Attention [27] utilized the attention mechanism [26] to model the importance of interactions. PeekFuture [15] additionally modeled the person-scene and person-object interaction with visual contexts. STGCNN [17] built a spatio-temporal graph of the scene with edge weights as the relative distance. SGCN [23] further imposed sparsity constrains on interactions and prune non-influential ones. M2I [25] classified agent relations with heuristics. However, they either implicitly learned the interactions or assumed fixed relations without considering the dynamics. We will consider the dynamic interactions by ranking their potentials temporally. Group detection was extensively studied [16, 21, 24] as a supervised classification task. However, these approaches become inadequate when handling datasets that lack group labels.

Contrastive learning with different modalities. CLIP [19] model shows contrastive learning effective in large-scale visual concept pre-training. Extended tasks of contrastive learning include object detection [22], text-image retrieval [4], and text-image segmentation [28], etc. In this study, we build a movement pattern embedding space in which the distance of historical and future patterns of a same person is minimized. We formulate our Contrastive History-Future learning as unsupervised pre-training.

3 Our Approach

We begin by outlining the notations and definitions associated with trajectory prediction tasks, followed by our approach description.

Notations. Let T_h be number of historical steps, and T_f be subsequent future steps. For a scene with N persons, their 2-D coordinates are denoted as $\boldsymbol{X}^h$ in history and $\boldsymbol{X}^f$ in future, respectively, as:

$$\begin{aligned} \boldsymbol{X}^h &= \{\boldsymbol{X}_i^h\}_{i=1}^N \ , \ s.t. \ \boldsymbol{X}_i^h = \{(x_i^t, y_i^t)\}_{t=1}^{T_h} \ ; \\ \boldsymbol{X}^f &= \{\boldsymbol{X}_i^f\}_{i=1}^N \ , \ s.t. \ \boldsymbol{X}_i^f = \{(x_i^t, y_i^t)\}_{t=T_h+1}^{T_h+T_f} \ . \end{aligned} \tag{1}$$

Let $\boldsymbol{G}$ denote a trajectory prediction model. $\boldsymbol{G}$ takes the N-person coordinates $\mathbf{X}^h$ as global context of the history and predicts the bi-Gaussian positional parameters (e.g., mean and variance of the XY-coordinates) for T_f future steps such as

$$\begin{aligned} \boldsymbol{Z} &= \{\boldsymbol{z}_i \in \mathcal{R}^{T_f \times 5}\}_{i=1}^N \ , \\ \text{with } \ \boldsymbol{z}_i &= \{(\mu_{x,i}^t, \mu_{y,i}^t, \sigma_{x,i}^t, \sigma_{y,i}^t, \rho_i^t)\}_{t=T_h+1}^{T_h+T_f} \ . \end{aligned} \tag{2}$$

For each person i of total N persons in the scene, we extract its D-dim historical feature $\boldsymbol{h}_i$ and its decoded future feature $\boldsymbol{f}_i$ as motion embeddings as:

$$\boldsymbol{H} = \{\boldsymbol{h}_i \in \mathcal{R}^D\}_{i=1}^N \ , \ \ \boldsymbol{F} = \{\boldsymbol{f}_j \in \mathcal{R}^D\}_{j=1}^N \ . \tag{3}$$

The collective outputs of $\boldsymbol{G}$ by Eq. 2 and 3 include the parameterized future predictions $\boldsymbol{Z}$, historical features $\boldsymbol{H}$ and future features $\boldsymbol{F}$, as:

$$\boldsymbol{Z}, \boldsymbol{H}, \boldsymbol{F} \leftarrow \boldsymbol{G}(\boldsymbol{X}^h) \ . \tag{4}$$

Depending on the backbone chosen for $\boldsymbol{G}$, $\boldsymbol{H}$ and $\boldsymbol{F}$ can be adaptively collected from the last (or pooled) hidden output of an LSTM, GCN, or Transformer.

3.1 Contrastive History-Prediction Learning

We introduce our Contrastive HIstory-Prediction (CHIP) learning to ensure the internal motion consistency of human behaviours. With historical motion embedding $\boldsymbol{H}$ and future embedding $\boldsymbol{F}$ of Eq. (3), we compute the dot-product for each (i, j) person pair such as $\boldsymbol{Q} = \{q_{ij} = \boldsymbol{h}_i \cdot \boldsymbol{f}_j\}_{i,j=1}^N$.

The concept of internal motion consistency dictates that the expected movement sequence of an individual, denoted as person i, ought to bear greater resemblance to the actual observed pattern compared to the rest of the individuals within the environment.

Consequently, every value along the diagonal of the matrix, represented by q_{ii}, must possess a higher magnitude compared to other entries in the corresponding row and column. Formally, for $i = 1, ..., N$, we have

$$q_{i,i} > q_{i,j} \ \wedge \ q_{i,i} > q_{k,i}, \ \ \forall j, k \neq i \ . \tag{5}$$

To impose above constraints, we formulate our CHIP learning objective as an auxiliary classification task such as

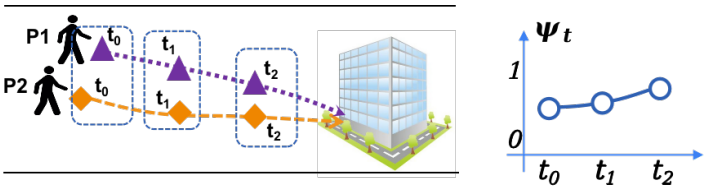
$$L^{chip}(\boldsymbol{Q}) = -\frac{1}{2N} \sum_{i=1}^{N} \big(\log \frac{e^{q_{ii}}}{\sum_{j=1}^{N} e^{q_{ij}}} + \log \frac{e^{q_{ii}}}{\sum_{k=1}^{N} e^{q_{ki}}} \big) \ , \tag{6}$$

in which q_{ii} gets maximized as a logit. As $\boldsymbol{Q}$ depends on feature embeddings from model outputs, we can optimize the model by minimizing L^{chip} with standard SGD.

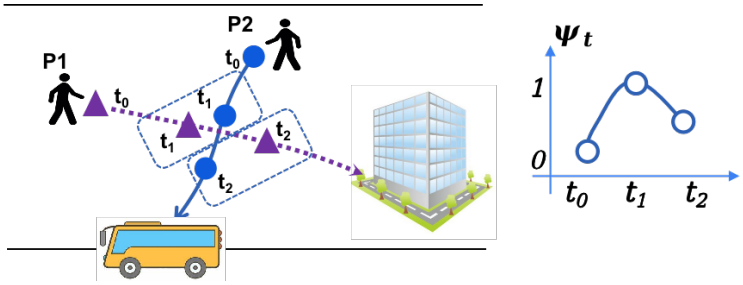
Notably, CHIP learning is unsupervised and can serve as a multi-task objective in model training. We describe the details in Sec. 3.4.

3.2 Differentiable Social Interaction Ranking

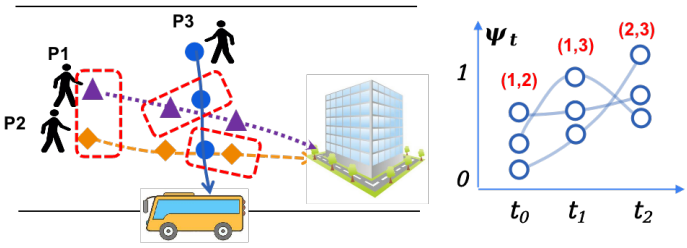
This study delves into explicitly modeling the properness of social interactions. To achieve this goal, we propose a novel Differentiable Social Interaction Ranking (DSIR) module, which enables the supervised optimization of predicted interactions among all participants to align with their actual positions.



(a) Walk together (in-group interaction).



(b) Walk and cross (out-group interaction).



(c) Interaction potentials (highest are highlighted).

Fig. 2: (a) In-group interaction with high potentials. (b) Out-group interactions with varied potentials at different steps. (c) Ranking of pairwise potentials. The interactions with highest potentials over the time are highlighted in dashed red box.

Observations. We highlight two important types of person interactions based on analysis of realistic data. Specifically, Fig. 2a illustrates the *in-group* interaction characterized by individuals walking in close proximity (i.e., less than 1 meter) as a same social group. Conversely, Fig. 2b captures the *out-group* interaction where two individuals walk independently but take measures to avoid collision by dynamically adjusting their walking paths while maintaining a comfortable social distance. This finding offers

insight into social dynamics and highlights the importance of considering various forms of interpersonal interactions in modeling pedestrian behavior.

We first define the *interaction intensity* of a paired persons (i, j) with the classical Gaussian potential function [2, 6, 29], depending on their relative distance d_{ij} as

$$\psi(i,j) = \exp\big(-\frac{d_{ij}^2}{2\sigma^2}\big) \in (0,1],\ \forall i \neq j\ , \tag{7}$$

in which σ is a constant social distance, d_{ij} is computed based on their positions, and ψ is symmetric ($\psi(i,j) = \psi(j,i)$). We omit $\psi(i,i)$ as we care about interactions with different persons. Thus, in a scene with $N \geq 0$ persons, we have $M = N(N-1)/2$ unique potentials to consider.

Social Descriptors. We propose to describe social interactions in a global view by pairwise potentials of pedestrians depending on their relative distances. The potential ranking over each step provides rich information about the temporal dynamics of human interactions. Fig. 2c shows a 3-person scene with complex interactions, i.e., P1 and P2 are in-group while P3 is out-group w.r.t to P1 and P2. Moreover, P3 is expected to interact with P1 and P2 by crossing their paths at subsequent steps. Thus, P1 and P2 need to plan their routes to maintain their intimacy while also keeping a polite distance from P3. As reflected in their ranks, $\psi(1,2)$ remains consistently high across all steps, while $\psi(1,3)$ and $\psi(2,3)$ reach their peaks at t_1 and t_2, respectively, before dropping off in later steps.

To further examine the complexity and appropriateness of human interactions, we propose to learn the predicted ranking of pairwise potentials using trajectory predictions and compare them with the actual order based on the ground truth data.

Task formulation. Given predicted pairwise potentials, we now formulate the task of ranking them increasingly as an optimization task with a differentiable solution. Thus we can integrate it into our learning procedure.

Let $\boldsymbol{\psi} = [\psi_1, ..., \psi_M]^\top \in \mathcal{R}^M$ be a list of M *unique* pairwise potentials in column form. Let $\mathcal{M} = \{1, 2, ..., M\}$ and $\mathbf{1}_M$ be an all-ones vector of dimension M.

We define an M-element *index array* $\boldsymbol{y}$ as

$$\boldsymbol{y} = [y_j = \frac{j}{M}]_{j=1}^{M} = \frac{1}{M}[1, \cdots, j, \cdots, M]^\top\ , \tag{8}$$

in which $y_j \in (0, 1]$ as similar as potential ψ, making the following sorting operations numerically stable.

Let $\boldsymbol{P} = \{p_{ij}\}_{i,j=1}^{M}$ be an $M \times M$ permutation matrix. $\boldsymbol{P}$ is binary in which each row or column only has one element of 1, otherwise 0. The 1-element p_{ij} ranks j-th element to i-th rank. A sorting permutation matrix $\boldsymbol{P^*}$ permutes $\boldsymbol{\psi}$ in increasing order as $\boldsymbol{\psi}^{P^*}$ as:

$$\boldsymbol{\psi}^{P^*} = \boldsymbol{P}^*\boldsymbol{\psi} = [\psi_1^{P^*}, ..., \psi_M^{P^*}]^\top,\ \forall i < j,\ \psi_i^{P^*} < \psi_j^{P^*}. \tag{9}$$

Sorting process is usually non-differentiable which requires comparing and swapping elements, e.g., QuickSort. We propose to formulate our potential ranking task as a differentiable learning objective and optimize it iteratively.

We first prove that a proper sorting permutation can be obtained by minimizing the following cost function.

Lemma 1. *Let $\boldsymbol{y} = [i/M]_{i=1}^{M}$ be the index array. Given a vector $\boldsymbol{\psi}$ of M unique elements, the sorting permutation P^* is the unique solution (out of all permutations) which minimizes the following cost*

$$L(\boldsymbol{\psi}^P|\boldsymbol{y}) = \sum_{i=1}^{M}(y_i - \psi_i^P)^2 \,. \tag{10}$$

We can perform proof by contradiction by supposing that there exists some non-sorting permutation P' ($P' \neq P^*$) which also minimizes Eq. (10). Then we can show if we further swap the non-sorted pairs we can further lower the cost.

Lemma 1 shows that sorting can be formulated as finding the optimal permutation that minimizes Eq. (10). Based on this, we construct a cost matrix $\boldsymbol{C}_{\boldsymbol{\psi}\boldsymbol{y}}$ as

$$\boldsymbol{C}_{\boldsymbol{\psi}\boldsymbol{y}} = \left\{c_{ij} = (y_j - \psi_i)^2\right\}_{i,j=1}^{M} \in \mathcal{R}^{M\times M}. \tag{11}$$

We formulate sorting operation as a relaxed integer programming as follows:

$$\begin{aligned} \hat{\boldsymbol{P}}^* &= \arg\min \langle \boldsymbol{P}, \boldsymbol{C}_{\boldsymbol{\psi}\boldsymbol{y}} \rangle - \lambda H(\boldsymbol{P}) \,, \\ s.t.\ \ \boldsymbol{P} &\geq 0 \,,\ \ \boldsymbol{P}\mathbf{1}_M = \mathbf{1}_M \,,\ \ \boldsymbol{P}^\top \mathbf{1}_M = \mathbf{1}_M \,, \end{aligned} \tag{12}$$

in which $H(\boldsymbol{P}) = -\sum_{i,j} P_{i,j} \log P_{i,j}$ is the entropy term.

The constraints of Problem (12) only limit each row and column of $\boldsymbol{P}$ sums to one and be positive, relaxing the requirement of a binary permutation matrix. This allows soft assignment of ranks, e.g., P_{ij} is interpreted as the weight of assigning element ψ_j to i-th rank.

The solution $\hat{\boldsymbol{P}}^*$ of Problem (12) can be solved in iterative and differentiable way [3].

Lemma 2. *For an $M \times M$ cost matrix $\boldsymbol{C}$, solving Problem (12) is strictly convex such that there exists a unique minimizer $\boldsymbol{P}^*$ which has the form of $\boldsymbol{P}^* = \boldsymbol{X}\boldsymbol{A}\boldsymbol{Y}$, where $\boldsymbol{A} = exp(-\lambda \boldsymbol{C})$ while $X, Y \in \mathcal{R}_+^{M\times M}$ are both non-negative diagonal matrices which are unique up to a multiplicative factor [3] , which can be efficiently solved with the differentiable Sinkhorn algorithm [3].*

Pairwise Ranking Loss. We can estimate the ranks of pairwise potentials with Task 12 as $\hat{R}(\boldsymbol{\psi}) = \{\hat{r}_i\}_{i=1}^{M} = M \cdot \hat{\boldsymbol{P}}^{*\top}\boldsymbol{y}$. In training stage, we can obtain the actual pairwise potentials $\boldsymbol{\phi}$ based on true positions $\boldsymbol{X}^f$, and sort to get their actual ranks $R(\boldsymbol{\phi}) = \{r_i\}_{i=1}^{M}$ as ground truths.

By comparing the predictions with the ground truths, we develop the social ranking loss L^{dsir} to penalize inconsistent pairwise orders in a supervised manner such that

$$L^{dsir}(\hat{R}|R) = \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\max(0, -(r_i - r_j)(\hat{r}_i - \hat{r}_j)) \,. \tag{13}$$

3.3 Bi-Gaussian Regression Loss

We follow previous studies [1, 8, 9, 23] to optimize the predicted trajectories with the bi-Gaussian distribution loss. Let the true future positions be $\boldsymbol{X}^f = (x, y)$ and the

parameterized model predictions be $\boldsymbol{Z} = (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ as in Eq. (2). We omit subscript i and t for simplicity. The bi-Gaussian distribution loss follows

$$L^{biG}(\boldsymbol{Z}|\boldsymbol{X}^f) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\Big(\frac{-1}{2(1-\rho^2)} \\ [\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}]\Big) , \tag{14}$$

which penalizes deviations of predicted (μ_x, μ_y) from the ground-truth (x, y) as well as large variances.

3.4 Two-Stage Multi-task Training Objective

In summary, we can optimize the model by jointly minimizing the bi-Gaussian regression loss in Eq. (14), CHIP learning loss in Eq. (6) and DSIR loss in Eq. (13) as

$$L^{final} = L^{biG} + \alpha_1 L^{chip} + \alpha_2 L^{dsir} , \tag{15}$$

in which α_1, α_2 are scaling factors. Specially, we propose a two-stage best practice of end-to-end model training with the multi-task objective L^{final} with standard SGD.

In Stage-1 *(Pre-training)*, we minimize $L^{pre} = L^{biG} + \alpha_1 L^{chip}$ for fast convergence, i.e., omitting DSIR loss. In Stage-2 *(Fine-tuning)*, we use the full L^{final} in Eq. (15) for fine-tuning trajectory predictions with interaction-aware DSIR loss.

4 Evaluation Metrics

We describe two standard metrics for trajectory prediction, then propose our novel *Intimacy-Politeness Score* to fully evaluate the properness of social interactions.

4.1 Standard ADE and FDE

ADE and FDE are two standard error metrics which measure the deviations from predicted positions to the ground truths. Let $(x_{i,t}, y_{i,t})$ be real position of person i at step t, and $(\hat{x}_{i,t}, \hat{y}_{i,t})$ be the predicted position. Their $L2$-distance is defined as $e_i^t = \sqrt{(\hat{x}_i^t - x_i^t)^2 + (\hat{y}_i^t - y_i^t)^2}$. The *Average Displacement Error* (ADE) [18] calculates the $L2$-distance between predicted future trajectory and ground truth, averaged over all future steps and all N persons in the scene as $\frac{1}{N \cdot T_f} \sum_{i=1}^{N} \sum_{t=T_h+1}^{T_h+T_f} e_i^t$. The *Final Displacement Error* (FDE) [1] computes the $L2$-distance between the predicted position and actual position at the final step as $\frac{1}{N} \sum_{i=1}^{N} e_i^{T_h+T_f}$.

4.2 Surrogate Social Distance Accuracy (SDA)

Public datasets often do not contain labels for social groups, as annotating social interactions can be a time-consuming process.

To address this issue, we suggest using a Social Distance Accuracy (SDA) scoring function to evaluate the quality of social interactions in a weakly-supervised manner.

The SDA method uses the pedestrians' predicted distances and compares them with weakly annotated group labels that are based on ground-truth distances. Our research demonstrates that SDA serves as a reliable surrogate measure for assessing the accuracy of trajectory predictions.

Let σ be the minimal distance of social politeness. Let $d_{i,j,t}$ be the actual relative distance between individuals i and j at a given time t, and let $\hat{d}_{i,j,t}$ be the predicted distance. We construct an adaptive in-group distance upper-bound $d^{+}_{i,j,t}$ and out-group distance lower-bound $d^{-}_{i,j,t}$, based on $d_{i,j,t}$ and σ. These two bounds define the acceptable social distance range in unsupervised manner without realistic group labels.

We choose a social distance threshold σ as the minimal distance of social politeness. Then we can establish the in-group person triplet set $\mathcal{D}^{In}$ and out-group person triplet set $\mathcal{D}^{Out}$ in unsupervised manner as follows:

$$\begin{aligned} \mathcal{D}^{In} &= \left\{(i,j,t)\middle|\bar{d}_{i,j,t} \leq \sigma \wedge d_{i,j,t} \leq \sigma\right\} , \\ \mathcal{D}^{Out} &= \left\{(i,j,t)\middle|\bar{d}_{i,j,t} > \sigma \wedge d_{i,j,t} > \sigma\right\} . \end{aligned} \tag{16}$$

We construct an adaptive range $[d^{-}_{i,j,t}, d^{+}_{i,j,t}]$ based on the actual distance $d_{i,j,t}$ with $\tau \in (0,1)$ such that

$$d^{+}_{i,j,t} = (1+\tau)\cdot d_{i,j,t} \text{ and } d^{-}_{i,j,t} = (1-\tau)\cdot d_{i,j,t} . \tag{17}$$

Now we design the hinge score functions s^I and s^P to measure intimacy and politeness separately, as follows:

$$\begin{aligned} s^{I}_{i,j,t} &= \textit{trunc}\left(\frac{d^{+}_{i,j,t} - \hat{d}_{i,j,t}}{d^{+}_{i,j,t} - d_{i,j,t}}, 0, 1\right), \forall (i,j,t) \in \mathcal{D}^{In} , \\ s^{P}_{i,j,t} &= \textit{trunc}\left(\frac{\hat{d}_{i,j,t} - d^{-}_{i,j,t}}{d_{i,j,t} - d^{-}_{i,j,t}}, 0, 1\right), \forall (i,j,t) \in \mathcal{D}^{Out} , \end{aligned} \tag{18}$$

in which $\textit{trunc}(\cdot, 0, 1)$ clips the value within $[0, 1]$. Concretely, s^I rewards a predicted $\hat{d}_{i,j,t}$ to be smaller than $d^{+}_{i,j,t}$ for in-group triplets in $\mathcal{D}^{In}$, while s^P rewards $\hat{d}_{i,j,t}$ to be larger than $d^{-}_{i,j,t}$ for out-group triplets in $\mathcal{D}^{Out}$.

Based on the social range, we design a hinge score functions s^I and s^P to measure intimacy and politeness separately, as Eq. 16 shows. In summary, S^I lessens when overestimating in-group distance d_1, while S^P lessens when under-estimating out-group distance d_2. We use the combined S^I and S^P as the Social Distance Accuracy (SDA).

5 Experiments

We first introduce datasets and benchmark models of existing works. Then we report the performance and carry out ablation studies to show the effectiveness of our methods.

Datasets. We use two widely compared public pedestrian trajectory datasets, i.e., ETH [18] and UCY [14] to evaluate our methods. In particular, ETH dataset contains the ETH and HOTEL scenes, while the UCY dataset contains the UNIV, ZARA1, and ZARA2 scenes. Each data sequence contains observed trajectories extracted from 8 frames (3.2 seconds) and future trajectories in the next 12 frames (4.8 seconds). The

train/val/test splits are given. Following a standard testing procedure [1,17], we generate 20 random samples from the predicted distribution for each testing trajectory, then we calculate the minimum ADE and FDE from the predictions to the ground truth. We also calculate the maximum SDA from all samples as our interactive metric value.

Table 1: Min ADE and FDE results on the benchmark ETH and UCY datasets.

Model	Architecture	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Vanilla-LSTM [1]	LSTM	1.09/2.41	0.86/1.91	0.61/1.31	0.41/0.88	0.52/1.11	0.70/1.52
Social-LSTM [1]	LSTM-Pool	1.09/2.35	0.79/1.76	0.67/1.40	0.47/1.00	0.56/1.17	0.72/1.54
Social-GAN [9]	GAN-Pool	0.87/1.62	0.67/1.37	0.76/1.52	0.35/0.68	0.42/0.84	0.61/1.21
Sophie [20]	GAN-Att	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.51/1.15
Social-BiGAT [11]	GAN-Att	0.68/1.29	0.68/1.40	0.57/1.29	0.29/0.60	0.37/0.75	0.52/1.07
STGCNN [17]	GCN	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
SGCN [23]	GCN-Att	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
SGCN+CHIP	Ours	0.59/0.92	0.29/0.51	0.38/0.72	0.28/0.49	0.25/0.45	0.36/0.62
SGCN+CHIP+DSIR	Ours	0.55/0.86	0.28/0.44	0.37/0.69	0.27/0.46	0.23/0.42	**0.34/0.58**

5.1 Our approaches and baselines

We construct our learning models based on the state-of-the-art SGCN [23]. **SGCN-CHIP** adds our proposed CHIP learning module to SGCN as a multi-task objective for model regularization. **SGCN-CHIP-DSIR** utilizes both CHIP and DSIR modules to boost training in a two-stage manner as in Sec. 3.4.

We compare our method with various existing methods such as Vanilla LSTM [10], Social-LSTM [1], Social-GAN [9], Sophie [20], Social-BiGAT [12], STGCNN [17] and SGCN [23].

5.2 ADE and FDE results

We show results evaluated with ADE and FDE (lower is better) in Table 1 and observe the following trends.

- The best performing SGCN-CHIP-DSIR leads SGCN by 8.1% in ADE (0.34 vs. 0.37) and 10.8% (0.58 vs. 0.65) in FDE relatively, establishing new benchmarks for pedestrian trajectory prediction.
- With only CHIP (no DSIR), our SGCN-CHIP still outperforms SGCN by 2.7% in ADE and 4.6% in FDE, showing the effectiveness of contrastive learning for regularizing the model training.
- By additionally performing DSIR, the SGCN-CHIP-DSIR outperforms SGCN-CHIP by 5.6% in ADE (0.34 vs. 0.36) and 6.5% (0.58 vs. 0.62) in FDE, relatively. This shows the effectiveness of potential ranking to explore interactive patterns.

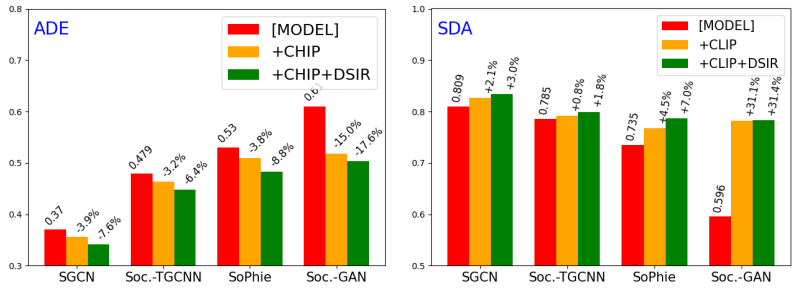
Table 2: SDA results. The higher is the better.

Model	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Van.-LSTM [1]	0.456	0.505	0.536	0.460	0.454	0.482
Soc.-LSTM [1]	0.463	0.497	0.544	0.461	0.450	0.483
Soc.-GAN [9]	0.522	0.542	0.581	0.606	0.640	0.578
Sophie [20]	0.692	0.683	0.740	0.735	0.825	0.735
STGCNN [17]	0.732	0.851	0.679	0.875	0.789	0.785
SGCN [23]	0.783	0.848	0.712	0.873	0.831	0.809
SGCN-CHIP	0.816	0.852	**0.723**	0.886	0.854	0.826
+DSIR	**0.832**	**0.854**	0.721	**0.891**	**0.870**	**0.834**

5.3 SDA results

We show results of SDA (higher is better) in Table 2 and observe the following trends. The best performing SGCN-CHIP-DSIR outperforms original SGCN by 3.1% in SDA (0.834 v.s. 0.809). Without DSIR, our SGCN-CHIP still leads original SGCN 2.1% over original SGCN (0.826 v.s. 0.809). STGCNN and SGCN have higher SDA with GCN backbones. Social-LSTM/-GAN have low SDA due to the ineffective social-pooling. Vanilla-LSTM has lowest SDA as it ignores interaction learning.

5.4 Ablation study of model adaptivity



(a) ADE (lower is better) (b) SDA (higher is better)

Fig. 3: Ablation studies of model compatibility.

We study whether our CHIP and DSIR learning modules can adapt to existing models and boost performance boost upon their original designs. We experiment with several recent studies: Social-GAN [9], Sophie [20], STGCNN [17] and SGCN [23]. They cover most existing backbones and social learning methods, including LSTM-GAN-Pool, LSTM-GAN-Attention, Transformer, GCN, and GCN-Attention, respectively, as shown in Table 1.

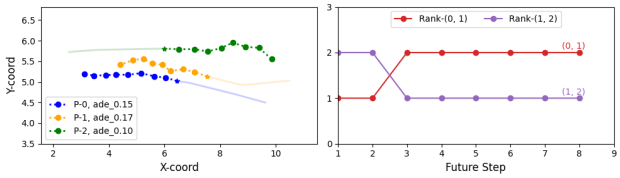
For each existing model, we build two variants as {MODEL}-CHIP and {MODEL}-CHIP-DSIR with one or both our proposed modules respectively. We show the performance boost in ADE and SDA in Fig. 3.

Fig. 3a shows that for each model except Social-GAN, our CHIP module lowers ADE by about 3-4%, and CHIP+DSIR together lower ADE by about 6-9%, relative to the original model. Our modules could improve Social-GAN by a substantial 15-18%, indicating the importance of regularizing the path generation process with GANs.

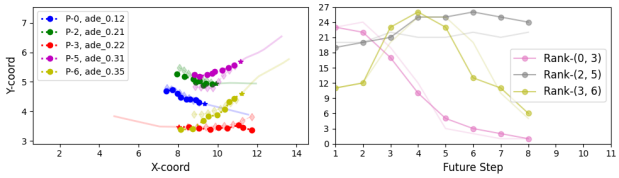
Fig. 3b shows that for each model except Social-GAN, our CHIP module can improve SDA about 0.8-4.5%, and CHIP+DSIR together can improve SDA about 1.8-7%. The improvement on Social-GAN as large as 31%.

Compare Fig. 3a and Fig. 3b, we found that SDA and ADE are generally consistent over most methods. This indicates that SDA serves as a trustful evaluation metric for unsupervised social interaction.

In conclusion, our model-agnostic learning modules can integrate into existing models and possibly future state-of-the-art models to achieve performance boosts readily.



(a) A 3-person scene (left) and pairwise ranking (right).



(b) A multi-group complex scene (left) and pairwise ranking (right).

Fig. 4: Visualize scene-level potential ranking.

5.5 Visualizations of Ranking

In Fig. 4a, we show a 3-person scene on the left and display their potential ranking over 8 future steps on the right. We have excluded pair P-$(0, 2)$ due to their significant distance. During steps 1 and 2, Rank-$(1, 2)$ is predominant since P-1 (yellow) and P-2 (green) are in closer proximity with higher potential. However, in step 3, P-2 moves away while P-1 continues to follow P-0 (blue) at a closer distance, which leads to a higher rank for P-$(0, 1)$. Fig. 4b illustrates a scene comprised of five individuals, as well

as the potential ranking of three pairs. On the left-hand side, the actual paths of each person are visually represented, while the right-hand side displays the corresponding potential rankings. The shaded lines in both figures show the actual ranking. As P-0 and P-3 move away from each other, the Rank-$(0, 3)$ drops over time. On the other hand, Rank-$(2, 5)$ remains at the top as P-2 and P-5 approach one another. Rank-$(3, 6)$ peaks at step-4, when P-3 and P-6 stand at the closest point before parting ways. Ultimately, our predictions align with the ground truths, demonstrating their accuracy.

6 Conclusion

This study introduces a framework that is model-agnostic and utilizes contrastive learning to ensure motion consistency and potential ranking for tracking interactions. Additionally, a novel metric has been developed to accurately quantify the interactive properness. The performance of our framework surpasses baselines by a significant margin, and can be seamlessly integrated into existing prediction models to enhance their performance. In future work, we aim to extend our methods to dense vehicle traffic scenarios where interactions between cars are constant, but less random due to the stricter constraints of traffic rules. Ultimately, the proposed methods can be integrated into self-driving technology as a crucial module for collision avoidance and path planning.

7 Acknowledgments

This work is supported by the National Natural Science Foundation of China, Project 62106156, and Starting Fund of South China Normal University.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: CVPR (2016)
2. Boykov, Y., Veksler, O., Zabih, R.: Markov random fields with efficient approximations. In: CVPR (1998)
3. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. In: NeurIPS. pp. 2292–2300 (2013)
4. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: CVPRW (2021)
5. Fang, L., Jiang, Q., Shi, J., Zhou, B.: Tpnet: Trajectory proposal network for motion prediction. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
6. Fathi, A., Hodgins, J.K., Rehg, J.M.: Social interactions: A first-person perspective. In: CVPR (2012)
7. Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., Schmid, C.: Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In: CVPR (2020)
8. Graves, A.: Generating sequences with recurrent neural networks. ArXiv **abs/1308.0850** (2013)
9. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: CVPR (2018)

10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**, 1735–80 (1997)
11. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I., Rezatofighi, S.H., Savarese, S.: Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In: NeurIPS (2019)
12. Kosaraju, V., Sadeghian, A., Martín-Martín, R., Reid, I.D., Rezatofighi, H., Savarese, S.: Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In: NIPS (2019)
13. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H.S., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: CVPR (2017)
14. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. Computer Graphics Forum (2007)
15. 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: CVPRW (2019)
16. Manfredi, M., Vezzani, R., Calderara, S., Cucchiara, R.: Detection of static groups and crowds gathered in open spaces by texture classification. Pattern Recognition Letters **44**, 39–48 (2014)
17. Mohamed, A., Qian, K., Elhoseiny, M., Claudel, C.: Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: CVPR (2020)
18. Pellegrini, S., Ess, A., Schindler, K., van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
19. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
20. 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: CVPR (2019)
21. Setti, F., Cristani, M.: Evaluating the group detection performance: The grode metrics. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(3), 566–580 (2019)
22. Shi, H., Hayat, M., Wu, Y., Cai, J.: Proposalclip: Unsupervised open-category object proposal generation via exploiting clip cues. In: CVPR (2022)
23. Shi, L., Wang, L., Long, C., Zhou, S., Zhou, M., Niu, Z., Hua, G.: Sparse graph convolution network for pedestrian trajectory prediction. In: CVPR (2021)
24. Solera, F., Calderara, S., Cucchiara, R.: Structured learning for detection of social groups in crowd. In: AVSS (2013)
25. Sun, Q., Huang, X., Gu, J., Williams, B.C., Zhao, H.: M2i: From factored marginal trajectory prediction to interactive prediction. In: CVPR (2022)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
27. Vemula, A., Muelling, K., Oh, J.: Social attention: Modeling attention in human crowds. In: ICRA (2018)
28. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: CVPR (2022)
29. Weiss, Y., Freeman, W.: On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Transactions on Information Theory **47**(2), 736–744 (2001)
30. Yu, C., Ma, X., Ren, J., Zhao, H., Yi, S.: Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In: ECCV (2020)
31. Zhang, P., Ouyang, W., Zhang, P., Xue, J., Zheng, N.: Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In: CVPR (2019)
32. Zhao, H., et al.: TNT: Target-driven trajectory prediction. arXiv preprint arXiv:2008.08294 (2020)