Contrastive learning of graph encoder for accelerating pedestrian trajectory prediction training

Zonggui Yao | Jun Yu | Jiajun Ding

Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China.

Correspondence
Jun Yu, Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China.
Email: yujun@hdu.edu.cn

Abstract
In the area of pedestrian trajectory prediction, the hybrid structures of temporal feature extractor or spatial feature extractor have paved the way for the precise prediction model, and they are in larger and larger scale. Learning of specific feature encoding model not only influenced by the structure of the network, but also by the learning manners such as supervised learning and unsupervised learning. Previous works concentrated on more comprehensive encoders and more delicate designs of feature extractors. However, the mutual influence factors from the neighbour pedestrians associate with the distance to the centre pedestrian seldomly noticed. Most of the existed feature extractors in prediction models trained in the way of supervised learning other than unsupervised manners caused the problem that the extracted features are always handcrafted without the natural distinction of obscure situations. The graph contrastive accelerating encoder is proposed, which accelerates the pedestrian trajectory prediction training process of the state of the art method of spatio-temporal graph transformer networks. Employing the unsupervised contrastive learning process and the graph of neighbours representing distance affection of nearest and farthest pedestrian to the centre pedestrian, the graph contrastive accelerating encoder significantly shrinked the training time. Holding the final performance on to state of the art level, the proposed method let the lowest pedestrian trajectory prediction error show up in the obviously earlier training steps.

1 INTRODUCTION

Pedestrian trajectory in the real world is influenced by a major factor of spatial man to man interaction, which is a tangled feature to extract because of the reality that this type of interaction cannot be predicted easily [1]. From this point, plenty works driven by data oriented methods, i.e. supervised learning, which aims to reveal the hidden features that representing the relationship among the pedestrians, as well as along time sequences. In these works, pedestrian centred method considers each one of the pedestrians from the aspect of target driven prediction [2], whereas relation centred method probe into crowd behaviour’s pattern and regard each frame of the target video as the object of spacial information extraction [3]. Some work additionally take time sequences as the research subject [4], to explore the interactions among the pedestrians in time sequences. Previous works always intended to deploy handcrafted feature extraction encoders such as graph neural networks (GNN) [5], long short-term memory (LSTM) [6], both LSTM and GNN [7], recurrent neural network (RNN) [8] and transformer model [1] to discriminate different state of pedestrian crowds. Most of these models ignored to figure out the reasonable route for training, and they assume the supervised learning is the proper training method or they seldomly consider the similarity measurement for training process.

The learning method always defines the performances of the designed encoders [9]. In most situations, the task does not need to figure out every detailed information for the reason that we only need the final destination of distinguishing among states, not to describe them. Furthermore, the unsupervised algorithms that can distinguish among tangly objects have better ability of encoding to generalize to unseen scenarios [10]. Recently, unsupervised learning achieved the comparable performance in tremendous areas of the tasks which originally
fulfilled by supervised learning [11]. Contrastive unsupervised learning is the most significant method in the unsupervised learning area, and the encoding of the object is suitable to employ contrastive learning [12]. This type of algorithm dealing with the encoding of images or situations has the original intention that handle the objective function from the perspective of information science, which employed similarity description functions to separate between the positive samples and the negative samples [13]. Thus make the training of the model along with more reasonable similarity based course. As far as current researches, supervised learning and contrastive learning will extract different kinds of features that we can assume their performances are complementary to each other [14].

GNN have been proved to extract the intrinsic connected nodes of their structure associative characters [15]. The GNN's theoretical support of feature relations adapted plenty areas of feature engineering or information extraction [16]. It is also a significant component in pedestrian trajectory prediction research [17]. Most of previous work employed GNN whose original purpose are discovering how the interactive model of all neighbour pedestrians at one frame of the video impact the final route selection of single pedestrian [18]. It more or less released some of the constrains of interaction detection problems. In associate with time sequences, some of the works expanded the graph concept to the temporal dimension [19]. The pooling operation among the neighbours of their spatial features in the trajectory prediction area is also effective [20]. Also in the thought of graph-based method, coherent groups [21], curriculum learning [22], weakly supervised [23] methods are researched previously. But the analysis of the pedestrian based graph in generalities will not be the optimal manner that to utilize graph concept. They don't depend on realistic interpretable characteristic of the pedestrian trajectory prediction scenarios.

The existed works concentrated on the traditional feature extraction or deep neural network design, seldomly pay attention to the training methods of the objective function design. Then the overall training process will go along a simplex route of the handcrafted model and the supervised criteria. On this condition, the structure of the prediction models and the tuning of the hyperparameters would be paid much more attentions. Other than supervised loss functions, contrastive unsupervised loss functions are always needless to design the detailed learning criteria of supervised labels, otherwise they only concentrate on the similarity function distinguish between positive samples and negative samples. These type of learning methods always based on the knowledge in information science, so it has more theoretic support. Also the previous works that used GNN never take the neighbour pedestrians’ influences of the nearest and farthest distances, that make the GNN's feature extraction efficiency remains low level. The GNN can describe the graphic relationships in real world only when the inside relationships are exactly interpreted. We cannot implement the GNN on the feature nodes without explore the inside mechanism of a specific scenario. The pedestrian trajectories are influenced by the neighbour pedestrians in a unsymmetrical way. We need to observe the significant influence factors of these unsymmetrical relationships and implement the GNN accordingly.

In this paper, we proposed that with the unsupervised learning as the pre-training of the encoders, the pedestrian trajectory prediction model will enabled with complementary learning benefits and the initial parameters of original prediction model would acquire advantage status that supervised learning would take large number of steps to achieve. We also employed graph convolutional network (GCN) as the feature extraction model. Differently, we added the aggregation function of maximum and minimum to identify the farthest and the nearest pedestrian from the centre pedestrian as Figure 1 described. In this way, we achieved both the interpretable model and the interactive awareness in the pedestrian trajectory prediction task.

To sum up, our contribution can be described as follows:

1) We proposed a unsupervised contrastive training method for the pedestrian trajectory prediction encoder;
2) We employed GNN for embedding the neighbour pedestrians with the aggregation function of minimum and maximum;
3) We worked out a synergetic training process that can take advantage of contrastive learning of GNN encoder and accelerated the original training process by a large percentage.

The rest of the paper is arranged as follows. Section 2 describes the recently released works which are relative to our topics. Section 3 presents the overall architecture of the proposed GCAE algorithm. Section 4 explains the implement details of the proposed algorithm. Section 5 present the experimental evaluations and ablation studies. The final Section 6 make the statement of the conclusion of the paper.

2 RELATED WORKS

The topic of pedestrian trajectory prediction has been researched widely in recent years. To enable the prediction model with the awareness of interactive information among the
pedestrians, Haddad et al. proposed the graph of pedestrians that representing the spacial change during $\Delta t$ [25]. Xue et al. proposed a social pyramid structure dedicated to explore human social connectivities in the same scene. Most researchers tend to regard the trajectory prediction problem as a spacial-temporal feature extraction framework. In [27] Zhang et al. employed classical LSTM networks, strengthen with the message passing mechanism among pedestrians. Xue et al. also deployed three different LSTMs to capture single human, among human and whole scene information [28]. Manh et al. incorporates scene-LSTM and pedestrian movement LSTM, these two coupled LSTMs refined each other to fulfil the prediction task [29]. Bisagno et al. are aware of the related person such as friends or families always have the same pattern of moving, came up with the idea of clustering the model of LSTMs [30]. Yu et al. employed currently outstanding performed transformer model and proposed to use transformer for human relationship and temporal variation extraction network [1].

Contrastive learning is one of the most significant methods of unsupervised learning. The role of contrastive learning always be the training of the encoders. Arora et al. described the contrastive unsupervised learning for representation systematically [32]. Dai et al. proposed the image captioning task can be improved by contrastive learning for encouraging distinctiveness as well as maintaining quality of the generated captions [31]. Kipf et al. use contrastive learning to represent the world models from image based inputs [33]. Mnih et al. propose to learn the word embeddings based on training log-bilinear models with contrastive learning and turned out faster and more effective [34]. In [35] the authors capture useful information from high-dimensional representations with contrastive learning method. Jiang et al. maximized feature consistency under differently augmented views to enhance the performance of their model in the process of unsupervised pre-training of contrastive learning [36]. Leveraging contrastive learning to capture high-level information between different modalities in a compact feature embedding, Meyer et al. found that the contrastive learning can be regarded as latent information among different modalities [37].

GNN recently attracted more and more attentions for its highly comparability to the connected relationship in the real world. Shen et al. proposed employ the GNN in the person re-identification task that considers other nodes’ information as part of similarity estimation [38]. Wang et al. noticed that GNN structure can be used in the area of pedestrian trajectory prediction with an unsupervised manner [39]. Mohamed et al. proposed that the GCN can be used as a interactive modelling tool in pedestrian trajectory prediction [40]. Liu et al. proposed to use GNN as the person re-identification model within and outside of the frame [41]. Shen et al. proposed multi-receptive field GCN of body parts graphs to enhance single-shot pedestrian detection [42]. Zhang et al. proposed that the GNN can be used for evaluating social relationships in pedestrian trajectory prediction. Liu et al. proposed to employ GCN to extract the spatio-temporal relationships for detection of the intention of pedestrians [44]. Cadena et al. evaluate intentions of pedestrians in the city scene employing pose estimation and GCN [45]. Xue et al. proposed that dynamic graphs are employed to describe the interactive relationship among pedestrians [46].

3 | ARCHITECTURE OF GCAE ALGORITHM

Our GCAE algorithm consists of two different phases in a dynamic learning process as Figure 2. The two phases are contrastive learning phase and synergetically training phase. And afterwards the original model would preserve the evaluation performance on the test sets. Parameters in previous process of contrastive learning can be regarded as a product of pre-train. The core idea of GCAE is that we did not change the original architecture of the objective pedestrian trajectory prediction model. We only implement a parallel inference GNN encoder model to accelerate the learning, and the two parallel models will complementary enhance each other by a final decoder layer. In this work, we employed the base framework with the state-of-the-art framework of transformer based prediction model [1]. Based on this framework, we found our GNN encoder can capture the critical latent features during the earlier iterations of the training. Thus make the overall synthetic model converged in a faster course. Our implementation mechanism make it possible that our GNN encoder and contrastive learning can be flexibly incorporate with multiple predictive model and accelerate their performance.

3.1 | The graph neural network encoder

In this part, we describe the GNN based encoder that is interpretable for human interactive behaviours. Our intention is to work out an effective way to reveal the hidden relative connections among the pedestrians in the same frame. These connections are not symmetrical such that we cannot design the networks from the symmetrical way of aggregation.

In the theory of graph convolution of spectral graph convolutional operation, the graph Fourier transform $\hat{F}$ to a signal $x$ can be denoted as:

$$\hat{F} = U^T x$$

(1)

where $U$ is the matrix of eigenvectors of the graph Laplacian matrix, we denote $L$ as:

$$L = U^T A U$$

(2)

$L$ can be described as the expansion:

$$L = I_n - D_{\frac{1}{2}} A D_{\frac{1}{2}}$$

(3)

where $A$ is the affinity matrix defined in the neighbour pedestrians’ graph, $I_n$ is the identity matrix, and $D$ is a diagonal matrix of node degrees with each diagonal element. Every element in
D can be expressed as:

\[ D_{ii} = \Sigma_{j} A_{ij} \]  

(4)

The graph convolution equals to the inverse Fourier transform of the product of Fourier transforms. Let \( x \) be the feature, and let \( c \) be the graph convolution kernel, we have:

\[ c \cdot x = U((U^T c) \odot (U^T x)) \]  

(5)

We can regard \( \theta = U^T c \) as the parameters of the convolution kernel, and define \( \varphi = \text{diag}(\theta) \), then we have:

\[ \varphi \cdot x = U\varphi U^T x \]  

(6)

By mathematical simplification [47], \( \varphi \cdot x \) can be approximated by:

\[ \theta(I_n + D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}}) \]  

(7)

In this equation, \( I_n + D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} \) can be further transformed to: \( D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} \), where \( \tilde{A} = I_n + \tilde{A} \) and \( D = \Sigma \tilde{A}_{ij} \).

We define the single pedestrian of node \( p \) in the interactive graph structure. Then the graph of the pedestrians existed in one frame denoted as \( G \) for the overall representation of graphic feature. The proposed GNN employed the GCN algorithm [47] as the foundation of the graphic feature, which embedding the features of neighbours to one node as:

\[ b = f(D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} X W) \]  

(8)

where \( f \) is the activation function, \( X \) is the neighbour pedestrians’ location feature representation, \( b \) is the centre pedestrian location feature representation, \( W \) is the parameters of the weight, \( \tilde{A} = I_n + \tilde{A} \) and \( \tilde{D} = \Sigma \tilde{A} \), where \( \tilde{A} \) is the affinity matrix according to pedestrians’ neighbouring relationship, \( I_n \) is the identity matrix, and \( D \) is a diagonal matrix of node degrees with each diagonal element. Additionally, we define the overall graph feature embedding method with the modification that make the aggregation function to be the combination of maximum and minimum function. After this process, the final graph feature \( H_G \) would be:

\[ b_{\max} = \max_{pi \in (0 \sim N)} b_{pi} \]  

(9)

\[ b_{\min} = \min_{pi \in (0 \sim N)} b_{pi} \]  

(10)

\[ H_G = \text{cat}_{i=0 \sim N}(\text{cat}(b_{\max}, b_{\min})) \]  

(11)

where \( N \) is the number of pedestrians, \( b_{pi} \) is the local feature of neighbour pedestrian \( pi \), \( b_{\max} \) and \( b_{\min} \) are the aggregated features of neighbour pedestrians, \( \text{cat} \) is the concatenate operation of vectors. The reason that we choose maximum and minimum as the aggregation function is that in pedestrian trajectory decisions, one man’s walking route will mostly influenced by the people who is nearest to the man. But considering the man making decisions will also evaluate the current situation of where his or her location of the largest scope, we concatenate the farthest pedestrian’s location into final graph feature. The graph feature contains the maximum and the minimum influences toward every pedestrian. To fit the output of the GCN model to the contrastive learning algorithm, the \( H_G \) of overall graph feature has been flatten into a \( N \times \text{size(feature)} \)
The GNN encoder design detail. The components are embedding layer, aggregation function and the encoding layer. The embedding layer and the encoding layer have the activation function of ReLu and the structure of fully connections.

![Diagram of GNN encoder design](image)

**Figure 3** The GNN encoder design detail. The components are embedding layer, aggregation function and the encoding layer. The embedding layer and the encoding layer have the activation function of ReLu and the structure of fully connections.

dimension vector. Then in the synergetically training phase, it recovered to the original size.

The GNN encoder can be described as Figure 3. The neighbour pedestrians of the centre pedestrian have their own location information. Then it is embedded by a one layer fully connected network and activated by a ReLu function to avoid the gradients to disappear. The embedded neighbour pedestrians' features need a dimensionality reduction for that the neighbours of centre pedestrian cannot be taken into consideration altogether. We need to pick up the most significant ones to construct the neighbouring situations. From this point, we used the aggregation function of maximum and minimum functionality.

The reason of doing so is that we believe that the surrounding pedestrians influence the centre pedestrian in a way that is not averaged. The influences of neighbour pedestrians should be arranged by the order of the distance to the centre pedestrian. The nearest neighbour pedestrians need to have the highest attention of the centre pedestrian. Another thinking of our model is that if we only used the minimum functionality as the aggregation function, then it cannot distinguish the situations that have multiple minimum feature nodes. Then we come up with the idea that the farthest pedestrians provide the centre pedestrian with the awareness of the whole scale of the free walking zone. It is also important for a pedestrian to know where can he or she go, that is some place that other people already there and it is the farthest people. After the aggregation process, then the aggregated features are fed into the final encoder of fully connected network with ReLU function to activate too. We add this layer for that we noticed if the number of parameters of an encoder is not enough, then it cannot be converged to the optimal status. So the final structure including some components that are not existed in common GCNs.

With the description above, the original pedestrian location feature will be embedded by one layer fully connected network, and the final graph embedding will also be encoded by one layer fully connected network. Then our proposed GNN encoder can be represented as:

\[
X' = \text{ReLU}(fc(X))
\]  

(12)

\[
\beta' = f(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X'W)
\]  

(13)

\[
b_{i \text{max}} = \max_{i=0}^{N} \beta_{i}'
\]  

(14)

\[
b_{i \text{min}} = \min_{i=0}^{N} \beta_{i}'
\]  

(15)

\[
H_G = \text{cat}(\beta_{i=0}^{N})(\text{ReLU}(fc(\text{cat}(b_{i \text{max}}, b_{i \text{min}}))))
\]  

(16)

where \(X'\) is the embedded local feature, \(\beta'\) is the graph embedded and local embedded feature, ReLU and \(fc\) are the activation function of ReLU and fully connected network respectively.

### 3.2 Contrastive training of GNN encoder

Graph encoder described in the previous part can be trained by either unsupervised or supervised method. We follow the course of unsupervised pre-train and directly inference in training process. The contrastive learning is performed during the pre-train phase, and the training is performed during the synergetic training phase. In the end, our proposed encoder will serve as the accelerator that can flexibly integrate into the SOTA prediction model. The construction of the GNN encoder is presented in the previous part. We denote the encoder as:

\[
\epsilon_{\text{GNN}} = \text{GNN}(l_{p1}, l_{p2}, ..., l_{pN})
\]  

(17)

where \(l_{p}\) are the locations of pedestrians in one frame, \(\epsilon_{\text{GNN}}\) is the encoding result. The encoding operation only exerted on each frame of the video sequences contains pedestrians. The spacial relationships are actually extracted twice by the synergetic model. First time is the original transformer prediction model, second time is our GNN encoder model. And the temporal information is extracted by the original transformer model. Our GNN encoder can be regarded as a spacial relationship prediction model to enhance the original transformer
model exerted on the spacial features. Also our GNN encoder can be regarded as a model that take the temporal impacts into consideration. It captured the most significant influences for a pedestrian and it will cause the future decisions of the pedestrian of the walking route.

After the encoding operation, the situation in one frame has been encoded into a \( N \times T \times \text{size(feature)} \) dimension vector. All the feature vectors in every frame built up the vector space of neighbour pedestrian trajectories. To measure the similarity in the context of the frame vector space is the necessary step for the next contrastive learning. In contrastive learning, the similarity is the key measurement for the model to distinguish positive samples and negative samples. It is the theory of information science that if the encoder can shrink the distance between original features and positive samples, meanwhile enlarge the distance between original features and negative samples, then the encoder model can do effective encoding.

The commonly implemented unsupervised contrastive learning methods dedicated to maximize the MI (mutual information) of the original features and its encoded latent features with joint density \( p(x, y) \) and marginal densities \( p(x) \) and \( p(y) \). The primitive MI objective function can be denoted as the expectation of:

\[
I(X, Y) = E_{p(x,y)} \left( \log \left( \frac{p(x, y)}{p(x) \cdot p(y)} \right) \right)
\]

(18)

where \( I(X, Y) \) is the mutual information. Consider all the measurement standards of independence between random variables, mutual information is outstanding because it has the information theoretic background. Not like the linear correlation coefficient, it is sensitive also to dependences which do not manifest themselves in the covariance [49]. We can notice that MI is zero if and only if the two random variables are strictly independent. If we consider the KL divergence, then the mutual information is the KL divergence:

\[
I(X, Y) = \text{KL}(p(x,y) \| p(x) \cdot p(y))
\]

(19)

We implement the training objective function as the directly optimization of the proportion of the gap between positive pairs and negative pairs for the mutual information estimation. Our contrastive loss estimator is:

\[
L_{\text{con}} = E \left( -\log \left( \frac{\exp(s(\text{pos}))}{\exp(s(\text{pos})) + \exp(s(\text{neg}))} \right) \right)
\]

(20)

where \( s(\text{pos}) \) is the similarity criteria between original features and minute deviation from original features, \( s(\text{neg}) \) is the similarity criteria between original features and large deviation from original features. In the context of pedestrian trajectory prediction, the \( s(\text{pos}) \) and the \( s(\text{pos}) \) is the similarity measurement. We need a proper measurement to describe the similarity between the original features and positive samples or negative samples. We consider the pedestrian trajectory prediction problem needs an predictive model that can extract the directional latent information of pedestrian’s trajectory. So we choose the cosine similarity function. It is defined as following function:

\[
s = \frac{\langle \mathbf{g}_{\text{GNN}} \cdot \mathbf{g}_{\text{GNN}}' \rangle}{\| \mathbf{g}_{\text{GNN}} \| \cdot \| \mathbf{g}_{\text{GNN}}' \|}
\]

(21)

where \( \langle \mathbf{g}_{\text{GNN}} \rangle \) is the deviation of the original features. This calculation is known as the cosine measure of the vector space. We employed this calculation for that the pedestrian decision making can be simplified as a directional policy, then the encoding method of the pedestrian trajectory should consider the latent similarity measurement as directional measurement of the vector space.

The generation method of positive samples and negative samples is the random transformation of the location coordinates of the pedestrians. The greater of the transformation scale, the closer from the samples to the negative samples. As a result, the generation of the positive samples and the negative samples can be described as following functions:

\[
D_{\text{positive}} = \mathbf{g}_{\text{GNN}} + \mu \times T_{\text{noise}}
\]

(22)

\[
D_{\text{negative}} = \mathbf{g}_{\text{GNN}} + \epsilon \times T_{\text{noise}}
\]

(23)

where \( \mu \) and \( \epsilon \) are the threshold of the noisy deviation to generate the positive samples and the negative samples. \( T_{\text{noise}} \) is the noise generation function that can output the noise within the range of \((-1, 1)\). \( D_{\text{positive}} \) and \( D_{\text{negative}} \) are the positive samples and negative samples accordingly.

### 3.3 Synergetic training processes

This part describes the working process of how the proposed encoder and contrastive learning synergetically trained with the base framework. The base of our work is the SOTA performance of transformer structure of spatial-temporal pedestrian trajectory prediction network [1]. This work consist of temporal process and spatial process. They firstly deployed the temporal transformer to extract temporal features within pedestrians, then the spatial extraction used TGConv which is a transformer-based message passing graph convolution mechanism. They built two encoders and each of them including a pair of spatial and temporal transformers, they stack the pair of spatial and temporal transformers to extract spatio-temporal information. Worth to notice, they used the standard transformer structure in most current works in which the core idea of is to replace the recurrence completely by multi-head self-attention mechanism. The contribution of our work is to accelerate the learning process of the original network and proposed an accelerate model that has the out-of-the-box functionality to adapt the other prediction network. The significant feature of our algorithm is that it never decreased the performance of the original network at all, meanwhile enhanced the convergence efficiency of the prediction model.
In the previous phase, the original framework of prediction model is ignored. The object of pre-training is only our proposed GNN encoder. The training method is the contrastive learning mentioned above. When the contrastive loss function that contains cosine measure of similarity started to converge, it means the encoder already begin to discriminate the negative samples form the positive ones. The goal of this phase is to let the GNN encoder has its own ability of extracting the latent features of the pedestrians’ relative information without any supervised labels. This is the major advantage of the proposed acceleration algorithm that needless of the pre-process of the dataset.

The second phase after the contrastive training is to add the GNN encoder into the original prediction model, and then synergistically train the synthetic model. The GNN encoder will be mounted to the layer before the decoder model of the original model, that is to mount before the fully connected layers after transformer model in our case. Because of the outstanding performance of the proposed encoder, the gradient of the overall network will be guided in much quicker course. It can be regarded as the well designed destabilization of the gradient for jumping out of the local optimum of the prediction function in much quicker course. In this way, the value of loss function of supervised synergetical training dropped down in very few iterations. Our synergetic training process take advantage of both model of the original transformer prediction model and the GNN encoder model. Then they synthesize in the final decoder layer. The weight distribution of each model will take the processed most effective features for the next stage of decoding. In this way, the effectiveness of the whole synthetic model enhanced. Actually, the advantages of supervised learning and unsupervised learning enhanced each other as well. There are multiple differences between supervised learning convergence course and unsupervised learning convergence course. They have their own advantage of convergence speed. The last decoder layer make it to let the fastest convergence course of each model part take in the learning process.

Taking full advantage of the dynamic computation graph, the overall convergence steps distinctly shrunk, and the original best performance never decreased. The flexibility of our proposed work process also can be migrated to any other prediction model for acceleration. The complete process of training is summarized in Algorithm 1. The data preprocessing follow the procedure that the coordinates of different pedestrians would be firstly arranged frame by frame, and in each frame, we marked all neighbour pedestrians of each pedestrian so that we can recognize the neighbours to formulate the graph. We also evaluated the real time performance on the test set during every iteration of training. This method make the evaluation of the acceleration in a fair way for the model is tested in real time to verify whether it is useful in actual applications.

The algorithm firstly take in the prepared dataset to the prediction model, then it initialized the parameters of the GNN encoder. In the contrastive learning phase, it processed the dataset and generated the positive samples and the negative samples in realtime of the contrastive learning process. In our case, we transformed the original coordinates of the pedestrians in the scales of 0.1 and 10 to generate the positive samples and negative samples respectively. The contrastive learning process takes place towards the objective function of equation 20 to shrink the distance between original features and positive features, enlarge the distance between the original features and the negative samples. Finally the contrastive learning enhanced the discrimination ability of the GNN encoder. After the contrastive learning of the GNN encoder, we freeze the parameters of the GNN encoder. This process is necessary because the synergetic training does not need to train the GNN encoder itself. The GNN encoder exists for its characteristics acquired from the contrastive learning process. It only serves as the accelerator of our synthetic model. So we only need the encoding ability of it to capture the latent features which the original model never revealed in the earlier training iterations. We mount the GNN encoder to the decoder of the original prediction model. Thus we can take advantages of transformer prediction model and the GNN encoder model. The synergetic training of the synthetic model goes towards the objective function of the original transformer prediction model, which is the supervised criteria towards the ground truth of the pedestrian trajectories. We also implemented the realtime evaluation on the test set to get the realtime application performance. The final evaluation is employed to verify the overall performance of the synthetic model.

## 4 IMPLEMENT DETAILS OF GCAE ALGORITHM

The implement of the GNN following the GCN model and modified some of the details. The node features are the pedestrian coordinates in a frame, which are the x and y coordinates of two dimensional trajectory. The neighbour pedestrians in one

---

**Algorithm 1 GCAE Algorithm**

**Data pre-process:** the coordinates of pedestrians are classified by frames and labelled by neighbour items.

**Input:** the preprocessed dataset of pedestrian trajectories.

**Initialization:** randomly initialize the parameters of the GNN encoder model, prediction model.

1: while not converged do
2:   take in the preprocessed dataset without the label,
3:   generate the positive samples and the negative samples,
4:   contrastive training of GNN encoder,
5: end while
6: freeze the parameters of the GNN encoder,
7: mount the GNN encoder into the original transformer model,
8: while iteration ≠ iter_max do
9:   synergistically training of the synthetic model,
10:   evaluation on test set along with the training,
11: end while
12: perform final evaluation.
frame constructed the connections of the pedestrian graph. All neighbours of one pedestrian connects to the centre pedestrian. Afterwards, the embed of one layer of fully connections of the coordinates expanded the channels into 16 for each one of the pedestrian. Then the aggregation functions of maximum and minimum achieved the dimensionality reduction from the shape of $N \times \text{size(feature)}$ into $1 \times 16$. Then we concatenated the maximum and minimum results of different channels into a single vector of 32 dimensions. It represents the surrounding situation of one pedestrian of nearest neighbour influence and farthest scale for space awareness. Then the latent features extracted above feed into a synthetic layer of full connected with ReLu activation function. This layer's input size is 32 channels and output size is 16 channels. In the end of the encoder, the latent features is in the shape of 16 per pedestrian.

In the phase of contrastive learning, the learning rate is set exactly the same to the base network training of transformer prediction model. The encoded features from different pedestrians has been concatenated together into a single vector for overall graph feature of contrastive learning. To avoid the gradient explosion, we performed the data normalization of pedestrian location information. The contrastive learning implemented a batch based gradient descent algorithm, and training for five epochs already met our demands for discrimination ability in cosine similarity measure value of above 0.95 between positive samples and original features.

The synergetic learning phase exerted the GNN encoder after the final temporal transformer layer, before the decoder layer of the original framework. To mount the proposed encoder into the original model, we expanded the input channel of decoder of additional 16 channels for fitting the output of GNN encoder. Then the other configurations of the original model are preserved. After another few steps, the base model converged into its best performance.

5 | EXPERIMENTS AND RESULTS

To demonstrate the effectiveness of GNN encoder after contrastive learning, we designed several experiments to test the convergence and evaluation of our proposed model. Our model achieved the best average displacement error (ADE) and final displacement error (FDE) in each iteration. We make it holding the original prediction model's performance, whereas the convergence steps shranked obviously, and the ADE and FDE evaluation also achieved the best in the earlier epochs of the training. Thus we achieved the acceleration of the learning process. We believe this is the reasonable result when we employed the unsupervised learning, which can be regarded as a complementary learning method with the supervised learning. Our algorithm can take advantages of both training methods, and the GNN model which is also implemented in a new way of nearest and farthest relationships extraction. The experimental results shows our method give the prediction model the awareness of most significant route decision impact factor, the scale for pedestrians to move, as well as the latent graphic features which cannot be extracted by supervised learning.

The experiments runs on the device of NVIDIA GeForce RTX 2080Ti, and the datasets are the same as the base work of transformer prediction model of ETH (ETH and HOTEL) and UCY (ZARA1, ZARA2, and UNIV) datasets. The evaluation is arranged as follows:

1) The contrastive learning convergence associate with the different aggregations of GNN encoder.
2) The convergence acceleration of the proposed method compare with the base method.
3) The ADE and FDE achieved along with the iterations.
4) The ADE and FDE achieved along with the iterations.
5) The ADE and FDE achieved along with the iterations.

The experimental results including the following studies:

1) The contrastive learning study that presents the proposed GNN encoder's unsupervised contrastive pre-training process;
2) The study of the synergetic training phase that mounted the previously pre-trained GNN encoder to the base SOTA transformer prediction model.
3) The ablation study of the GNN encoder's different structures and the different mounted locations to the original model;
4) The discussion of different manners to improve the performances better based on the experimental results.

5.1 | Contrastive learning

Our first experiment is that we employed the contrastive learning of the GNN model, with datasets except of the labels. This type of unsupervised learning is the key to capture the latent information in a natural way, which needless of the handcrafted design of the criteria in training. The original datasets of pedestrian trajectory prediction need to be preprocessed in the order of frames and identify different pedestrians in the same frame, as well as identify the neighbours of single pedestrian. As the preprocess of the datasets finished, we formulated the graph of single pedestrian according to the neighbours distribution. Then the GNN is constructed accordingly.

The GNN is constructed in the following way. First, all the neighbour pedestrians of every pedestrians has been found in one frame of the pedestrian video. second, we fetch the coordinates of these neighbour pedestrians to be the nodes of the GNN network. Every nodes of the GNN network has the embedding layer of fully connected network. These networks share the parameters of the weights. Then the embedding products get into the dimensionality reduction operation. We implemented the aggregation function of maximum and minimum functionality to shrink the dimensions of the embedded latent feature as well as aggregate the graphic features. The aggregated features will go through another embedding process of one layer of fully connected network. Then the GNN encoder constructed completely.

For contrastive learning, We need to generate the positive samples and the negative samples for training. We performed
the coordinates transformation of the pedestrians, such that the positive samples are the ones with the minimum transformations, whereas the negative ones with the maximum transformations. In our experiments, we used the transformation threshold of 0.1 and 10 for positive and negative samples respectively. The contrastive training of the GNN realized with the loss function of equation 20. After the contrastive learning converged, the GNN encoder is enabled with the discrimination of different situations of graph of pedestrians. The evaluation of the cosine similarity between the original data and positive samples can achieve above 0.95. The reason that we choose the cosine similarity is the latent features represent the directional influence for a single pedestrian. We think the influence is the mental impact force. When a man meets the nearest man right in his or her way, then his or her decision should be influenced in the pedestrian's walking route and we think it formulated a decision influence force.

After the contrastive learning of the GNN encoder, the encoder has the outputs that encoded the maximum and minimum influences of the neighbour pedestrians. We visualized the influences from the neighbour pedestrians that extracted by the GNN encoder in Figure 4. The arrow in Figure 4 marked the major influence from the nearest neighbour, i.e. from the GNN encoded neighbour minimum feature distance. Actually we can regard the encoding as the interpretation of the most significant mental impact force to the centre pedestrian. It will be the major influence of the walking route. Another major influence is the farthest pedestrian, which decide how large of the scale of one pedestrian’s observation area, marked with blue polygon in Figure 4. It also plays an important role of the trajectory decision for that the trajectories always within the observation area which is always relative to the farthest pedestrian.

To design the feature extractor for the most significant neighbour features, we tested and evaluated different aggregation function for the best performance of the GNN encoder. As Table 1 shows, the best combination is the maximum and minimum aggregation function together and afterwards of the feature embeddings. That means, we need to embed the original features and then aggregate the neighbours’ embedded features at the corresponding channels to achieve the highest level of discrimination. The results also proved the reality that pedestrian trajectories are influenced significantly by the nearest neighbour pedestrian and the scale of free walk space associate to the farthest neighbour pedestrian. It accord with the fact that when a man is the nearest neighbour of one pedestrian, then the

| Aggregation function          | Contrastive loss |
|------------------------------|------------------|
| Aggregation max              | 0.42             |
| Aggregation min              | 0.41             |
| Aggregation sum              | 0.52             |
| Aggregation avg              | 0.58             |
| Aggregation ahead            | 0.43             |
| Aggregation afterwards       | **0.31**         |

Figure 4. The visualization of the two major influences for one pedestrian, nearest and farthest neighbour pedestrians, corresponding to the two aggregation functions of our GNN encoder. Pedestrian 1 is the centre pedestrian, pedestrian 2 and 3 are the nearest and farthest pedestrians respectively. Red arrow represents the route decision impact force to the centre pedestrian, yellow arrow represents the ground truth trajectory. Blue polygon describes the free walking zone of the centre pedestrian. The scene is from the dataset of HOTEL.
The aggregation function has been analysed in previous paper [48]. The maximum and minimum aggregation function has its own shortcomings in distinguishing ability. We specifically used the combination of the maximum and minimum functionality to give consideration to both of the aggregation methods. In this way when the max functionality fails in distinguishing multi-maximum situations, the minimum functionality can be a complementary distinguishing tool for different situations of pedestrians. The average aggregation functionality cannot converge to expected level, we think it is because of the nodes of pedestrians formulate a graph in the way of asymmetrical way. That is to say, the influences of neighbour pedestrians are not in an equal way. This encourages our idea that the pedestrians’ impacts to the centre pedestrian cannot be take in without filtrate. Sum functionality also treat every nodes of pedestrians equally. These symmetrical methods of aggregation will hide the useful information with the equal operations.

walking route of this pedestrian will significantly affected by him or her. Whereas the farthest neighbour pedestrian always gives the concept of the free walking zone. So far we just performed the operation in one single frame, we think when this operation generalized to the time sequences, the influences should not be simply add up of these effects mentioned above.

The experimental results of Tables 3 and 4 demonstrate that the accelerated model surpasses the original model of the SOTA model. In the earlier 170 epochs, GNN enhanced model has its own shortcomings in distinguishing ability. This table only presented the few epochs of fast dropping loss. When the training process goes to epoch 17, our synthetic model firstly achieved double zeros after the point. During the whole process of the earlier epochs, the proposed method always ahead of the original model in loss convergence. The fast drop of loss in few early epochs make the advantage of our algorithm much more obvious. Afterwards, the advantage is not so obvious and we think the convergence acceleration is already done. But when the loss convergence acceleration is done, we noticed that the evaluated scores of ADE and FDE on the test set continues to accelerate.

The synthetic model that concatenated the original SOTA model and our GNN encoder has the characteristic of the original modelling ability, enhanced with our faster convergence ability. Every time our model achieved the lowest loss value, the original model fall behind our model, without influencing the original performance of convergence. The gradients of the synthetic model pass through the original model and ours, the weights of our synthetic model adjusts themselves to the distribution that take the advantages of both model, for a faster convergence course of descent. The final performance of the prediction model defined by the handcrafted designs of details. Our GNN encoder cannot help with the final delicate designs of feature extraction, but can help with the gradient descent in a faster course. We think it is because that our model revealed some part of the relative information directly, but the overall interpretation ability is weaker than the original model. Our model can only be the accelerator other than the substitution of the original SOTA transformer prediction model.

Along with the loss convergence, the ADE and FDE also decreased faster. We evaluate the ADE and the FDE on the test set of hotel. The performance of the synergetic model always achieved the best performance of the ADE and FDE scores earlier than the original model. This indicated that the proposed encoder not only helped to accelerate the training process, it also helped to accelerate the evaluation performance of the original model along with the training epochs. We presented the epochs that the accelerated model surpasses the original model of the ADE and FDE associate with training epochs in Tables 3 and 4. And this experiment takes place on the test set of HOTEL.

The experimental results of Tables 3 and 4 demonstrate that our proposed synthetic model achieved faster training performance. We can tell from the obvious gap with the original SOTA model. In the earlier 170 epochs, GNN enhanced model achieved the best ADE and FDE performances faster all the

| Epoch | Loss original | Loss accelerated |
|-------|---------------|------------------|
| Zero shot | 4.38779 | 3.91550 |
| 0 | 0.97325 | 0.89469 |
| 1 | 0.16375 | 0.14861 |
| 2 | 0.06630 | 0.06294 |
| 4 | 0.03386 | 0.02598 |
| 5 | 0.02866 | 0.02010 |
| 15 | 0.01620 | 0.01272 |
| 17 | 0.01014 | 0.00974 |

| Epoch | ADE best original | ADE best accelerated |
|-------|------------------|----------------------|
| 55 | 0.694 | 0.280 |
| 105 | 0.291 | 0.268 |
| 115 | 0.291 | 0.200 |
| 170 | 0.291 | 0.197 |
time. The gap becomes much more greater in the situation of FDE performance. Even though the FDE performance has less statistical meanings than ADE, but the FDE demonstrate the final trajectory prediction accuracy of the a prediction model. We think these two evaluation methods should be regarded in a comprehensive way. We can also tell that the epochs around 55 are the periods that our model performs much better. We think it shows the ability of adaption of our model.

The ADE and FDE performances show us the evidences of the acceleration in convergence. In the evaluation process, the parameters of the overall model already have been frozen, and the model entered the evaluation state. We check the actual performance of a model along with the training process for that the convergence acceleration can be watched in another perspective. The accuracy of a prediction model cannot be judged by the loss convergence, but it can be judged by the evaluation accuracy in real time. We are excited to see that even the advantage in loss convergence is not so obvious, but the evaluation advantages are so obvious of real time ADE and FDE. The corresponding advantage can be calculated by percentage of 12.4 and 40 of ADE and FDE respectively. In this situation, we think it is the strong evidence of acceleration of the training process. Most of our experiments have been performed multiple times for the prediction model is a probabilistic model which has the noise generation in it. We can also see during the earlier epochs of the convergence, our model’s error already dropped to the finest performances of the overall model. This is also the strong evidence of acceleration in learning process.

Here we describe the convergence curve in the synergetic training phase on the dataset of HOTEL. As Figures 5 and 6 presented, the earlier 120 epochs of the convergence proved the synthetic model has surpassed the original model by an obvious advantage. This figure generated at the time step which the corresponding model achieved the best ADE or FDE performances, then the best performance hold until the next best performance shows up. We can tell from the figures that the ADE and FDE performances of our model achieved the bottom of the curve during the earlier epochs already. Also, this experiment takes place on the test set of HOTEL.

We can see in Figure 5 that, because of the model complexity, the ADE performance is not good at the first few iterations. But after only five iterations, the ADE has dropped to the lowest level that the original model can achieve after 100 iterations. We think this acceleration is an obvious advantage. As we can see in the figure, our best ADE achieved performance has a rise curve after some iterations, it is because we need to achieve both the best performance of ADE and FDE, so we updated the ADE and FDE whenever the best performance of any of them shows up. We can see after about 110 iterations, the original model could achieve the best performance that our model achieved at iteration 55. We think it is the strong evidence that we can accelerate the training process with the proposed method.

In Figure 6 we can see the same situation shown up in the earlier epochs. The model complexity caused the first several epochs of the proposed model has high FDE performance. But after epoch 55, the advantage of our model shown up. And the FDE performance of our model always lead the best FDE achieved curve. The same performance of FDE achieved after 90 epochs of the original model, whereas ours achieved this level in epoch 55. We can see our model’s curve has dropped so fast that achieved the model’s potential best FDE in short
time steps. And compare to ADE performance, the FDE performance has more stable advantage to the original model. Summarized from these two performance figure and consider the experimental results in Tables 3 and 4, we can see our algorithm has the advantages of faster convergence, faster achieved the ADE and FDE performance, and faster making it of the final best performance.

The first few iteration’s curve of ADE and FDE of the synthetic model is not converged to the low level of the original model because of the model complexity, but the proposed model can quickly converge to let the ADE and FDE value drop in only few number of iterations. We think it is another strong evidence of acceleration of the convergence. Because of the ADE and FDE values represent the actual performance of the model, we think the actual performance of the synthetic model converges to the optimal state in a faster course. In most of the iterations in our experiments, our model converged only in very few number of iterations, whereas the original model achieved the same level of our model taking much more training iterations. From multiple perspective, we always can tell that our model is the fastest model in learning process.

We evaluated the proposed algorithm on most of the common tested datasets of pedestrian trajectory prediction, Which are ETH, HOTEL, ZARA1, ZARA2, and UNIV. The training performance always accelerated by a large percentage in the earlier epochs. The results presented in Table 5. The acceleration performances are calculated in the first 120 epochs because after these epochs the performances of original model and our model are already stabled. The best performance appears on the dataset of HOTEL, and the worst performance appears on the dataset of ZARA2. We think the pedestrian density is the key that caused the differences between the datasets. We also can see from Table 5 that in FDE performance the proposed method can achieve better performance than the ADE performance. It mainly because of the ADE performance has more statistical meaning that not easily to be improved.

The acceleration of different datasets measured in the same way of the HOTEL dataset. Here we only presented the acceleration in percentage. The convergence of loss, ADE and FDE performance is similar to the HOTEL dataset mentioned above. We report the acceleration percentage to show our model’s generalization ability for different datasets. The presented performances are tested in multiple experiments and we captured the best performances on different datasets. Our model performances always accelerated the training process. On ZARA1 and ZARA2 datasets our model only accelerated the performances of ADE a little bit. We think the datasets have something in common that we can extract the common feature to achieve the acceleration. But the density of different datasets cannot be uniformed. We added some tricks to make the proposed model perform continually well such as add or decrease the channels of the proposed structure, enlarge or shrink the noise of the model. After all, the overall architecture of the proposed model never changed. We think in the future works we can explore the stable version of the GNN encoder that can be more effective to accelerate the learning process and even improve the final performances. So far we already know the inside mechanism of how the unsupervised trained GNN encoder help the prediction model in the pedestrian trajectory prediction task.

### 5.3 Ablation study

The ablation study including the different structures of the GNN encoder itself and the different mounted positions of the original model described in Tables 6 and 7 respectively. The general view for these two ablation study are Figures 7 and 8. For GNN encoder itself, the features from the datasets firstly get through an embedding layer that expands the original features of the 2D pedestrian location to 16 dimensions. We tested if we do not employ this layer, the performance will dramatically drop down. We think it is because that the influence forces to a specific pedestrian is multi-directional. Thus if we decompose the 2D spacial features to multi-directional features, it will gain the precision of the encoder model. We further tested if we don’t employ the GNN as the main structure of the encoder, then the encoder never helped to improve the performance of pedestrian trajectory prediction. It proved that our method of nearest and

| Dataset | ADE acceleration (%) | FDE acceleration (%) |
|---------|----------------------|----------------------|
| ETH     | 11.2 35              |
| HOTEL   | 12.4 40              |
| ZARA1   | 9.8 27               |
| ZARA2   | 8.8 28               |
| UNIV    | 10.2 21              |

| Structure | Mean-ADE acceleration (%) | Mean-FDE acceleration (%) |
|-----------|---------------------------|---------------------------|
| No embed  | 5.3 8.7                   |
| No graph  | −3.2 −5.1                |
| No aggregation | 2.3 1.8          |
| No encode layer | 4.6 7.7          |
| Altogether | 10.5 30.2               |

| Structure | Mean-ADE acceleration (%) | Mean-FDE acceleration (%) |
|-----------|---------------------------|---------------------------|
| GNN-direct | −2.3 −4.4              |
| GNN-spatial encoder | −6.4 −8.9        |
| GNN-temporal encoder | −10.4 −17.3       |
| GNN-decoder | 10.5 30.2              |
The ablation study that implement different structures of the GNN encoder. The ablation situations are the GNN encoder without original feature embedding, the GNN encoder without graph of the neighbour pedestrians, the GNN encoder without aggregation methods that taking in all of the embedded features, the GNN encoder without the final encoding layer. Our model implemented all of the components above.

Farthest neighbour thinking is reasonable and adapts the task of pedestrian trajectory prediction well. The aggregation function ablation already has been tested at the beginning of the Experiment Section, here we tested if we remove the aggregation part of the encoder, that is to concatenate all the graph embed nodes’ features directly. This test operation wiped out our dimensionality reduction operation, as well as the intention of emphasis on the nearest and farthest pedestrian features. The performance of no aggregation model barely improved the overall performance. In the end of the GNN and aggregation operation, we deployed one layer of fully connected encoder. Turned out it enhanced the adaption ability that meets the contrastive learning feature encoding. The GNN encoder that integrates the components mentioned above altogether achieved the best acceleration performance.

From Table 6 we can notice that except for the proposed architecture of GNN encoder, other structures barely achieved the performance acceleration. The most significant component is the graph structure of our model. We think the graph model and its detailed structures such as the aggregation mechanism are the effective part of our model. We employed the nearest and farthest aggregation mechanism that can give the model the awareness of who is the most significant pedestrian to take into consideration. And we can also judge from the experimental result that the embedding and the encoding operations are also necessary components for a graph encoder. Altogether our GNN encoder utilized all the useful part which we tested and improved final performances.

The contrastive loss describes the evaluation standard of the cosine similarity for a specific structure of GNN encoder. The more positive samples getting closer to the original features, and the negative samples getting farther to the original samples, the better of the GNN encoder of its discrimination ability to identify different situations of pedestrians in one frame. We further discovered that if the embedding and the encode layers are not added into our model, the contrastive training convergence will stop at earlier epochs. We think this phenomenon is the evidence that the encoding of a graph shaped feature need the number of parameters to achieve to some level, otherwise the interpretation ability will be weaken.

In the next ablation study we tried different mounted locations to the original model. These methods are employing GNN encoder as one of the serial encoders directly of overall prediction model, putting GNN outputs into spacial transformer input, into temporal transformer input, into the final decoder of the original prediction model. As far as we evaluated, in Table 7, the method that putting GNN outputs into the final decoder of the original prediction model is the most effective way and performs the best acceleration results.

As we can see in Table 7, if we mount the GNN encoder as a part of serial encoders and preprocessed the data for spatial encoder, then it is not working. We think it is the problem of the combination manners. If we need to expand our model to the time sequences, we cannot simply stack the GNN model. Same problems appear in the situations of regarding the GNN encoder to be a part of spacial or temporal encoder. We think the transformer spacial encoder and the transformer temporal encoder need the primitive features as well such that they can always extract the useful information, not to put an encoder in the first place and hinder the abundant information from being extracted. This mechanism explains why the other combinations of the mounted locations fail to achieve the acceleration of learning. They blocked
FIGURE 8 The ablation study of where the output of GNN encoder mounted to different positions of the original transformer prediction model. The arrow of green colour represents the output of GNN encoder feed into spacial transformer. The arrow of purple colour represents the output of GNN encoder feed into temporal transformer. The arrow of orange colour represents the output of GNN encoder feed into the final decoder. Our method choose to let the output of GNN encoder feed into the final decoder.

TABLE 8 GPU hours advantage of GCAE algorithm

| Structure | GPU hours during ADE 0.7 | GPU hours during FDE 1.7 |
|-----------|-------------------------|--------------------------|
| Original  | 1.54 h                  | 2.87 h                   |
| Ours      | 1.25 h                  | 1.92 h                   |

the information flow so that the performances decreased even though we tried to fine tune the GNN encoder.

5.4 Discussion

Based on the experimental results above, we discuss the experiences after our experiments and the future improvements in this part.

We firstly present the GPU hours taken in our experiment. As we can see in Table 8, the GPU hours that our algorithm taken to achieve the same performance also improved significantly in the earlier training time. During the reaching of dramatically improving of the performances in ADE and FDE, no matter how the training iteration goes, we can see the GPU hours also shorter than the original model.

About the similarity measurement in unsupervised contrastive learning, we also studied it in multiple methods. The Euclidean distance is the most commonly used and easy to be used measurement. But it need all the dataset in the same scale and the problem of curse of dimensionality also make us give up on this measurement. The Manhattan distance works well in problems have binary system in it, but this measurement is not designed for the shortest distance, it means we cannot measure the shortest similarity in our case. The Chebyshev distance has the p value that we cannot easily know how to decide it, also we does not need the linear measurements. The Jaccard measurement is fit for the set operation and not suitable for our task.

Our algorithm of GCAE can accelerate the learning process but it cannot improve the final performance of the original model. We think it is because that the interpretation ability of our GNN encoder model is not strong enough to compare with the original transformer model. But we can also see the obvious acceleration performance of our GNN encoder. This phenomenon proved our model can extract the features which the original model cannot capture in the earlier learning epochs. The complementary learning process finally has the acceleration ability for training. The structure of our model need to expand the GNN encode operation to the time step dimension. But based on the experiments that we tried to do so, the directly stack operation on the time sequences of our GNN encoder never work. We think there should be the high dimensional operations which can go across the spacial and temporal differences. These methods of extract the features across spacial and temporal differences have been researched, but most of them regarded the spacial and temporal feature extraction at the same perspective. Thus their works cannot achieve the outstanding performance. We think if the dimensions of spacial and temporal features can be abstracted to a high dimensional latent feature and be considered together, then the performance would be much better for the high dimensional abstraction can regard the spacial and temporal features in the same way.

The graph based express of the real world also have been studied in many works, but we think the reason of the slow process for GNN based usage is that they seldomly see this problem in the perspective of the real mechanism of real relations. Take pedestrians interactions for example, a man will never take the pedestrians whose distance from himself or herself is too far into consideration. Actually, the nearest pedestrians’ trajectories are the most significant factors for a man to considerate. A walking people will dodge form the nearest pedestrians whom is in his or her way. Another factor is that a man need to be aware of his or her free moving space. This is another critical factor to influence one’s trajectory. People will not go to the area that
nobody else will go in common. These two factors are the base-
ment of our work. So the implementation of GNN will not work
until the mechanism adapt the real world. It need delicate obser-
vation of the real world to summarize the inside mechanism.
The hyperparameters of the GNN also need to be fine tuned
to adapt the task. For pedestrian trajectory prediction, the num-
ber of pedestrians in one frame is in the scale of about hundred.
If the GNN design is for a large scale graph embedding, then
the delicate details will be ignored. The dimensionality reduc-
tion function is the key to a GNN network to decide the scale
of the objective data.

The learning method defines the course to get the classifi-
cation or prediction expectation. different methods of conver-
gence such as supervised and unsupervised learning will affect
the final performance of converged model. For example, the
contrastive learning can be regarded as a learning method which
never rely on the handcrafted criteria such as MSE loss or
Crossentropy loss, but rely on the natural course of similarity
approach. If we train an encoder with the will of distinguish the
situations, then the handcrafted criteria is always insufficient for
a task. At least we can consider the supervised learning and the
unsupervised contrastive learning as two complementary meth-
ods for an efficient course of learning.

6 CONCLUSION
In this paper, we proposed an acceleration algorithm employed
GNN as the extra encoder, training with contrastive learning,
for the task of pedestrian trajectory prediction. Our method
takes advantages of the unsupervised learning of the encoder
to be an complementary module for prediction model. And
our algorithm employed the GNN as the pedestrian environ-
ment awareness model, that takes the significant nearest and
farthest neighbour pedestrians into consideration. We enabled
our GNN encoder with the awareness of obstacle awareness
and free walking space awareness. The proposed model synthet-
ically trained with the original SOTA transformer model, the
overall training process has been accelerated. The results has
been tested in multiple datasets and proved to be helpful for the
pedestrian trajectory prediction task. The evaluation best per-
fomrances of ADE and FDE achieved at earlier epochs of the
training process and the advantages are obvious. We believe that
with further works, this type of unsupervised contrastive learn-
ing will help more prediction models to fulfil the pedestrian tra-
jectory prediction task.

ORCID
Zonggui Yao https://orcid.org/0000-0003-1619-2732

REFERENCES
1. Yu, C., et al.: Spatio-temporal graph transformer networks for pedes-
trian trajectory prediction. In: Proceedings of the European Confer-
ence on Computer Vision Workshops, pp. 507–523. Springer, Berlin
(2020)
2. Zhao, H., et al.: TNT: Target-driven trajectory prediction. arXiv:2008.
8229v2 (2020)
3. Li, J., et al.: Social-WAGDAT: Interaction-aware trajectory prediction via
wasserstein graph double attention network. arXiv:2002.06241 (2020)
4. Huang, Y., et al.: STGAT: Modeling spatial-temporal interactions for
human trajectory prediction. In: Proceedings of the IEEE/CVF Interna-
tional Conference on Computer Vision, pp. 6271–6280. IEEE, Piscataway,
NJ (2019)
5. Haddad, S., Lam, S.-K.: Self-growing spatial graph network for context-
aware pedestrian trajectory prediction. arXiv:2012.06320 (2020)
6. Alahi, A., et al.: Social lstm: Human trajectory prediction in crowded
spaces. In: Proceedings of the IEEE conference on computer vision and
pattern recognition, pp. 961–971. IEEE, Piscataway, NJ (2016)
7. Sun, J., Jiang, Q., Lu, C.: Recursive social behavior graph for trajectory
prediction. In: Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition, pp. 660–669. IEEE, Piscataway, NJ (2020)
8. Saleh, K., Hossny, M., Nahavandi, S.: Intent prediction of pedestrians via
motion trajectories using stacked recurrent neural networks. IEEE Trans-
actions on Intelligent Vehicles 3(4), 414–424 (2018)
9. Love, B.C.: Comparing supervised and unsupervised category learning.
Psychonomic Bulletin & Review 9(4), 829–833 (2002)
10. Mirza, M., Courville, A., Bengio, Y.: Generalizable features from unsuper-
vised learning. arXiv:1612.03809 (2016)
11. Jainwal, A., et al.: A survey on contrastive self-supervised learning. Tech-
nologies 9(1), 2 (2021)
12. Le-Khac, P.H., Healy, G., Smeaton, A.F.: Contrastive representation learn-
ing: A framework and review. IEEE Access 8, 139397–139394 (2020)
13. Laskin, M., Srinivas, A., Abbeel, P.: CURL: Contrastive unsupervised rep-
resentations for reinforcement learning. In: International Conference on
Machine Learning, pp. 5639–5650. MLR Press, Boston, MA (2020)
14. Chen, J., Sinha, S., Kyriilidis, A.: IMCLR: Implicit contrastive learning for
image classification. arXiv:2011.12618 (2020)
15. Scarselli, F., et al.: The graph neural network model. IEEE Trans. Neural
Netw. 20(1), 61–80 (2008)
16. Fan, W., et al.: Graph neural networks for social recommendation. In: The
World Wide Web Conference, pp. 417–426. ACM Press, New York (2019)
17. Zhang, L., She, Q., Guo, P.: Stochastic trajectory prediction with social
graph network. arXiv preprint arXiv:1907.10233 (2019)
18. Monti, A., et al.: Dag-net: Double attentive graph neural network for tra-
jectory forecasting. arXiv:2005.12661 (2020)
19. Haddad, S., et al.: Situation-aware pedestrian trajectory prediction with
spatio-temporal attention model. arXiv:1902.05437 (2019)
20. Nikhil, N., Tran Morris, B.: Convolutional neural network for trajectory
prediction. In: Proceedings of the European Conference on Computer
Vision Workshops, pp. 186–196. Springer, Berlin (2018)
21. Wang, Q., et al.: Detecting coherent groups in crowd scenes by multiview
clustering. IEEE Trans. Pattern Anal. Mach. Intell. 42(1), 46–58 (2018)
22. Wang, Q., et al.: Density-aware curriculum learning for crowd count-
ing. IEEE Transactions on Cybernetics, https://doi.org/10.1109/TCYB.
2020.3033428 (2020)
23. Wang, Q., Gao, J., Li, X.: Weakly supervised adversarial domain adapta-
tion for semantic segmentation in urban scenes. IEEE Trans. Image Process.
28(9), 4376–4386 (2019)
24. Xu, Y., Piao, Z., Gao, S.: Encoding crowd interaction with deep neural
network for pedestrian trajectory prediction. In: Proceedings of the IEEE
Conference on Computer Vision and Pattern Recognition, pp. 5275–5284.
IEEE, Piscataway, NJ (2018)
25. Haddad, S., Lam, S.-K.: Self-growing spatial graph networks for pedestrian
trajectory prediction. In: Proceedings of the IEEE/CVF Winter Confer-
ence on Applications of Computer Vision, pp. 1140–1148. IEEE, Piscata-
way, NJ (2020)
26. Xue, H., Huynh, D.Q., Reynolds, M.: Pedestrian trajectory prediction using
a social pyramid. In: Pacific Rim International Conference on Artificial
Intelligence, pp. 439–453. Springer, Berlin (2019)
27. Zhang, P., et al.: SR-LSTM: State refinement for lstm towards pedestrian
trajectory prediction. In: Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition, pp. 12077–12086. IEEE, Piscata-
way, NJ (2020)
28. Xue, H., Huynh, D.Q., Reynolds, M.: SS-LSTM: A hierarchical lstm model
for pedestrian trajectory prediction. In: 2018 IEEE Winter Conference on
29. Manh, H., Alaghband, G.: Scene-LSTM: A model for human trajectory prediction. *arXiv:1808.04018* (2018)
30. Bisagno, N., Zhang, B., Conci, N.: Group LSTM: Group trajectory prediction in crowded scenarios. In: Proceedings of the European Conference on Computer Vision Workshops, pp. 213–225. Springer, Berlin (2018)
31. Dai, B., Lin, D.: Contrastive learning for image captioning. In: Advances in Neural Information Processing Systems, pp. 898–907. MIT Press, Boston, MA (2017)
32. Saunshi, N., et al.: A theoretical analysis of contrastive unsupervised representation learning. In: International Conference on Machine Learning, pp. 5628–5637. MLR Press, Boston, MA (2019)
33. Kipf, T., van der Pol, E., Welling, M.: Contrastive learning of structured world models. Paper presented at the 7th International Conference on Learning Representations, New Orleans, LA, 6–9 May 2019
34. Mnih, A., Kavukcuoglu, K.: Learning word embeddings efficiently with noise-contrastive estimation. *Advances in Neural Information Processing Systems* 26, 2265–2273 (2013)
35. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv:1807.03748* (2018)
36. Jiang, Z., et al.: Robust pre-training by adversarial contrastive learning. *arXiv:2010.13137* (2020)
37. Meyer, J., et al.: Improving unimodal object recognition with multimodal contrastive learning. In: IEEE/RSJ International Conference on Intelligent Robots and Systems 2(3), 4 (2020)
38. Shen, Y., et al.: Person re-identification with deep similarity-guided graph neural network. In: Proceedings of the European Conference on Computer Vision, pp. 486–504. Springer-Verlag, Berlin, Heidelberg (2018)
39. Mohamed, A., et al.: Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14412–14420. IEEE, Piscataway, NJ (2020)
40. Liu, G., Wu, J.: Video-based person re-identification by intra-frame and inter-frame graph neural network. *Image Vis. Comput.* 106, 104068 (2020)
41. Shen, C., et al.: Multi-receptive field graph convolutional neural networks for pedestrian detection. *IET Intelligent Transport Systems* 13(9), 1319–1328 (2019)
42. Zhang, H.: Generative pedestrian trajectory prediction with graph representation. *Journal of Physics: Conference Series* 1651(1), 012149 (2020)
43. Liu, B., et al.: Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters* 5(2), 3485–3492 (2020)
44. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv:1807.03748* (2018)
45. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. *arXiv:1609.02907* (2016)
46. Xu, K., et al.: How powerful are graph neural networks? Paper presented at the 6th International Conference on Learning Representations, Vancouver, 30 April–3 May 2018
47. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. *Phys. Rev. E* 69(6), 066138 (2004)