TrajCross: Trajecotry Cross-Modal Retrieval with Contrastive Learning

Quanliang Jing[1,2], Di Yao[1], Chang Gong[1,2], Xinxin Fan[1], Baoli Wang[1,2], Haining Tan[1,2], Jingping Bi[1]
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China[1]
University of Chinese Academy of Sciences, China[2]
{jingquanliang, yaodi, fanxinxin, wangbaoli, tanhaining, bjp}@ict.ac.cn

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671305

***Abstract*—In this paper, we propose a new task namely trajectory cross-modal retrieval which achieves the cross-modal search between coordinate trajectories and images containing trajectories. Nevertheless, trajectory cross-modal retrieval is rather challenging in learning the representations of each modality and reduce the cross-domain discrepancy caused by the inconsistent data distribution at the same time. we proposes a cross-modal retrieval model TrajCross based on multi-level representation for trajectory cross-modal retrieval. Specifically, TrajCross extracts the location features and the shape information respectively for the represention of multi-modal data. we adopt a contrastive learning method to achieve semantic preservation among similar multi-modal data. Extensive experiments show that TrajCross significantly outperforms state-of-the-art cross-modal retrieval methods.**

***Index Terms*—Cross-modal retrieval, Trajectory, Iamge, Contrastive Learning**

I. Introduction

With the development of IoT techniques, massive trajectories are collected in different ways, such as GPS sensors, surveillance videos and *etc*. The modalities of these data are different which leads that the analysis result of each modality can not be shared. Cross-modal retrieval, aiming to achieve mutual retrieval between two or more modalities, builds a bridge across different modalities and is vital for trajectory data analysis.

However, trajectory cross-modal retrieval is challenging due to the multi-modalities and semantic gap as shown in Figure 1. on one hand, the data format and composition of multi-modal data are different. It bring the challenge in designing model to extract right information for cross-modal retrieval. On the other hand, the semantic information contained in the two modalities are different. Building the information extraction model for each modality separately can not ensure the right part to be extracted.

Various methods have been proposed for cross-modal retrieval but none of them is designed for trajectory data. Existing methods for cross-modal retrieval mainly use deep learning based methods. Most of them employ deep neural networks and build two sub-networks for different modalities to learn the characteristics of each modality separately. Then the two models are linked together with a joint layer to learn cross-modal correlation. Although these methods can model images, the characteristics of the trajectory data cannot be extracted well. Besides these cross-modal retrieval method, there exist other works designed to learn trajectory embeddings [15], [16]. However, these method can only model coordinate trajectories and not capable of modeling images. Above all, there has no method can be directly used for trajectory cross-modal retrieval.

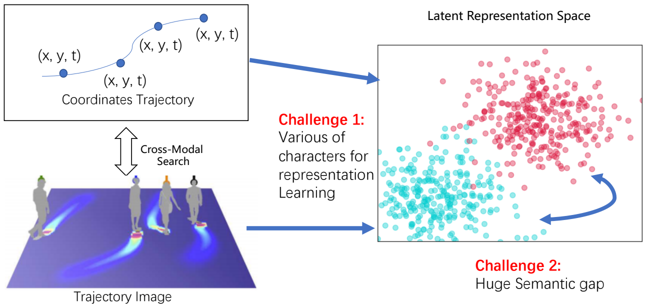

Fig. 1. An illustration of trajectory cross-modal retrieval.

To overcome the limitations, we propose a novel method, namely TrajCross, to systematically transform the semantic information of both the coordinate trajectories and trajectory images in one shared latent space. For the coordinates trajectories, TrajCross extracts the location features and the shape information respectively. As the auxiliary feature, the shape information is converted to an image to help the information extraction of image trajectories. Subsequently, we employ adversarial network to bridge the semantic gap of cross-modal representation.

To summarize, the contributions of our work are as follows:

- We define the trajectory cross-modal retrieval task. To the best of our knowledge, it's the first work trying to solve the trajectory cross-modal retrieval problem on coordinate trajectories and trajectory images.
- We propose a novel model, namely TrajCross, to systematicly learn the representation of different modalities of data ensuring these representations to be in the same latent space.
- Through extensive experiments on real-world datasets, we illustrate the effectiveness of the model in trajectory

Jingping Bi and Di Yao (bjp,yaodi@ict.ac.cn) are the corresponding authors.

978-1-6654-3902-2/21/$31.00 ©2021 IEEE

cross-modal retrieval.

II. Related Work

In this section, we briefly review the representative methods of cross-modal retrieval. The methods can be divided into two categories: traditional statistical analysis methods and deep learning based methods.

Traditional statistical analysis methods. These methods mainly learn linear projection matrices through traditional statistical analysis methods. It projects the characteristics into the public space and obtains a public representation. Canonical Correlation Analysis (CCA) [6] is one of the most representative works. The solution achieves retrieval by maximizing the correlation between cross-modal datasets. In addition, many methods combine label information in order to improve the performance of the CCA. For example, the work [12] incorporate semantic category label to improve the performance of CCA. This type of method mainly learns linear projection to maximize the correlation between paired data of different modalities.

Deep learning based methods. Due to the powerful ability of deep learning to model highly nonlinear correlations, various cross-modal retrieval methods based on deep learning have been proposed. To alleviate the problem of limited linear ability of CCA, a method DCCA [2] combining CCA and deep learning was proposed. DCCA maximizes the correlation at the top of the two subnets. Srivastava et al. [13] proposed a multimodal deep belief network (Multimodal DBN), which uses two different modal DBNs to model the distribution of the original features. In addition, methods based on likelihood analysis have also been proposed [8]. Although this kind of method based on deep learning can better characterize the correlation between multi-modality, the characteristics of the trajectory data cannot be extracted well, which will harm the retrieval accuracy.

III. THE PROPOSED MODEL

A. Preliminaries

There is a Dataset S which is a collection of N instances of trajectory-image pairs. A trajectory T_i is represented by a sequence of points collected from various equipment. The image I_i has the location information of its geographic area, and at the same time, the image contains the trajectory line. Formally,

$$S = \{T_i, I_i\}_{i=1}^{N},\ T_i = \{lat_i^k, lon_i^k\}_{k=1}^{M},$$

$$I_i^{loc} = \{U_i, B_i\},\ U_i = \{lat_i^u, lon_i^u\},\ B_i = \{lat_i^b, lon_i^b\}.$$

T_i and I_i are the corresponding trajectory and image of the i-th matched pair. lat_i^k and lon_i^k represent the longitude, latitude for trajectory T_i at timestamp k and M is the length of the trajectory. U_i and B_i are the latitude and longitude of the geographic location corresponding to the upper left and lower right corners of the image.

The goal of this paper is to achieve cross-modal retrieval of trajectories and images containing trajectories. Given a trajectory containing S points, we need to retrieve images that match the given S points from millions of images containing trajectories, and vice versa. The core of the problem is to encode the data of different modalities to get the embeddings, and to establish a measurement method $D(T_i, I_j)$ to judge the similarity between these embeddings.

B. Trajectory Representation

Previous works obtained the trajectory representation using recurrent neural networks. However, the extracted features have no physical meaning. We adopt a multi-feature representation for the trajectory. Specifically, we consider the position and shape of the trajectory sequence to obtain the embedding of the trajectory.

Position representation of the trajectory. We adopt the CNN method to extract the position representations of the trajectory sequences. Let $T = \{t_1, t_2, \cdots, t_m\}$ denote a trajectory with $t_m = \{lat_m, lng_m\}$, where $m \in \{1, 2, \cdots, M\}$ denotes the m-th timestamped point and M is the length of T. First, we use a linear layer to embed each point as multi-dimensional continuous vectors.

$$\mathbf{t_m^e} = TrajectoryEmbed(t_m). \tag{1}$$

Then, all the embeddings (row-wise) are aligned vertically and sequentially to form the trajectory matrices, denoted as $\mathbf{T_m^e} = \{\mathbf{t_1^e}, \mathbf{t_2^e}, \cdots, \mathbf{t_m^e}\}$.

In the convolution stage, convolution filters of different heights are applied to the matrix. Note that the width of the filter is the same as the length of the trajectory embedding. The output of the convolution is a vector with different lengths. But after the max pooling of each vector, it can be concated to form the initial representation of the trajectory, followed by fully-connected layers. The whole process is recorded as:

$$\mathbf{P_t^e} = LinearLayer(TrajectoryCNN(\mathbf{T_m^e})). \tag{2}$$

Shape representation of the trajectory. Different from the previous method of directly applying CNN to the trajectory sequence to obtain the trajectory position representation, we convert the trajectory sequence into an image in order to better obtain the shape information of the trajectory sequence based on the CNN, followed by a fully connected layer. We denote the converted image as $\mathbf{IMG_t}$ and the shape representation of the trajectory as $\mathbf{S_t^e}$:

$$\mathbf{S_t^e} = LinearLayer(TrajectoryImgCNN(\mathbf{IMG_t})), \tag{3}$$

After obtaining the shape representation and position representation of the trajectory sequence, we obtain the final representation of the trajectory sequence by concating them, denoted as $\mathbf{T^e}$

$$\mathbf{T^e} = concat(\mathbf{P_t^e}, \mathbf{S_t^e}). \tag{4}$$

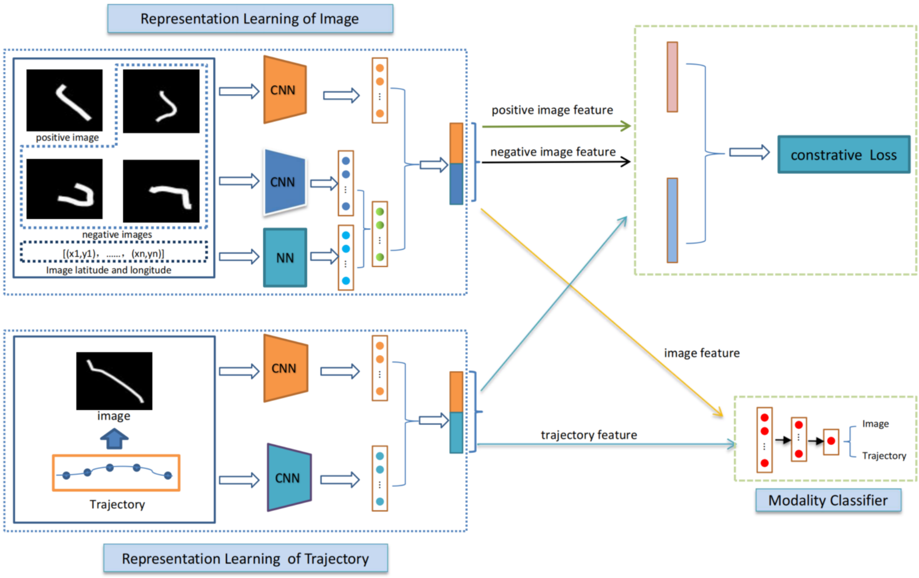

Fig. 2. Architecture of the TRAJCROSS. The function of the representation of trajectory and image is to obtain the representation of the trajectory sequence and image. Afterwards, the embeddings are mapped to the common space through adversarial learning to eliminate cross-modal discrepancy.

C. *Image Representation*

As with the image trajectory sequence, we also need to consider the representation of the image from the two aspects of position and shape.

Position representation of the image. How to learn the position of the trajectory in the image without prior knowledge is an important factor for the success of cross-modal retrieval. In order to obtain the position of the trajectory in the image, this paper uses an improved CNN method to process the image, denoted as WCNN. The core of the improvement is to make the width of the convolution kernel the same as the width of the image. The advantage of the improvement is that the convolution operation is not limited to small areas, but extract features from larger receptive fields. We denote the position representation of the Image as P_i^e

$$\mathbf{P_i^e} = LinearLayer(WCNN(\mathbf{IMG})), \tag{5}$$

where $\mathbf{IMG}$ represents the image corresponding to the trajectory sequence.

The above operation obtains the relative position of the trajectory in the image. We also need to determine the absolute position of the trajectory. First, we embed the location information of the image as $\mathbf{V^e_{locInfo}}$,

$$\mathbf{V^e_{locInfo}} = LinearLayer(\mathbf{I_i^{loc}}). \tag{6}$$

Then, in order to obtain the absolute position representation of the trajectory, we concat the representation of the image position $\mathbf{V^e_{locInfo}}$ and the representation of the trajectory position in the image $\mathbf{P_i^e}$, followed by a fully connected layer,

$$\mathbf{V^e_{locInfo}} = LinearLayer(concat(\mathbf{P_i^e}, \mathbf{V^e_{locInfo}})). \tag{7}$$

Shape representation of the image. As we discussed above, various CNN methods have been able to extract high-level semantic information, so we directly use the CNN method to obtain the shape semantic information of the trajectory from the image, denoted as:

$$\mathbf{S_i^e} = ImgCNN(\mathbf{IMG}), \tag{8}$$

We get the final representation of the image by concating the shape representation and position representation of the image, denoted as I:

$$\mathbf{I^e} = concat(\mathbf{P_i^e}, \mathbf{S_i^e}). \tag{9}$$

D. *Objective Formulation*

As illustrated in Fig 2, there are positive images and negative images. We defined the matching trajectory and image pairs as positive samples, represented here as $\{T, I_+\}$. We use images that do not match the trajectory to form a negative sample, denoted as $\{T, I_-\}$.

Adversarial loss. Due to the domain gap between trajectory sequence and image, the extracted features from both domains cannot be matched for similarity measurement. Similar to the work [17], this paper aims to learn a common subspace to enable cross-modal comparison. we adopt GAN to eliminate the gap between cross-modal.

We define a modal classifier, denoted as S, to distinguish whether the embedding representation comes from an image or a trajectory. The representation of the image is assigned label 0, and the representation from the trajectory is assigned label 1.

Constrative loss. The core of cross-modal retrieval is to measure the similarity between modalities. After obtaining the trajectory representation $\mathbf{T^e}$ and the image representation $\mathbf{I^e}$, the similarity between the image and trajectory is measured with a dot product, $(\mathbf{T^e})^{\mathbf{t}}\mathbf{I^e}$. In order to achieve cross-modal retrieval, we aim to minimize the gap between the representations of all semantically similar items from different modalities, while maximizing the distance between semantically different items from the same modality. To achieve this goal, we train the model by using the standard noise contrast estimation (NCE) [5], [10]:

$$\mathcal{L}_{con} = -\sum_{i=1}^{n} log \left(\frac{(\mathbf{T_i^e})^{\mathbf{t}}\mathbf{I_i^e}}{(\mathbf{T_i^e})^{\mathbf{t}}\mathbf{I_i^e} + \Phi} \right), \qquad (10)$$

$$\Phi = \sum_{I_- \in \psi_i} (\mathbf{T_i^e})^{\mathbf{t}}\mathbf{I_-^e}, \qquad (11)$$

where ψ_i represents the negative sample set of the image relative to the trajectory. It contrasts the score of the positive pair to a set of negative pairs sampled from a negative set.

Overall loss. The modules of TrajCross are trained end-to-end. However, the parameters of modules are optimized separately using different loss functions based on the above. The process of training the model runs as a minimax game, and the goals of the two loss functions are opposite. The process of learning the best embedding is conducted by jointly minimizing adversarial loss and embedding loss.

IV. Experiments

All experiments will be implemented in PyTorch 1.0 on a workstation with GPU NVIDIA 2080ti and Ubuntu operating system.

Datasets. Our experiments are conducted on simulated datasets. The dataset is generated based on univ [11], which has been widely used by many other works [1], [4]. Following the data processing principle as reported in [4], the position of the pedestrian from the original pixel locations in the image will be converted to world coordinates in meters. We convert each trajectory into an image. These generated images are used as the image to be retrieved in cross-modal retrieval. In order to make the generated image more practical in line with the actual situation, we will randomly generate the thickness of the trajectory line in the image when generating the image corresponding to the trajectory. Finally, We generated 10,000 training data and 2379 testing data. This dataset serves as a benchmark for the performance evaluatione on our proposed model.

Evaluation Metric. Since there is only one ground-truth match for each trajectory/image, to evaluate the retrieval performance of the proposed schemes, we use the measure of top N percentage, which is a classical performance evaluation criterion in the research on cross-modal retrieval for datasets without semantic label [3], [9]. More specifically, an trajectory(image) is considered correctly retrieved if it appears in the first k list created from its corresponding image(trajectory). Similar to [14], two cross-modal retrieval tasks are considered for the proposed mode: trajectory retrieval from an image query, and image retrieval from a trajectory query.

Compared models. To verify the effectiveness of our model, we compare our approach with state-of-the-art methods, which have been widely adopted as benchmarks in other literatures. It is notable that all the methods to be compared do not consider location information. To make a fair comparison, we add the location information to the embedding of the image when implementing the comparison method.

- **CCA** [6]. This is an important technique to learn a common subspace for heterogeneous data. It establishes the correlation between two sets of data through a linear analysis method.
- **MOCO** [7]. This paper uses contrast learning to learn modal features. After we convert the trajectory into an image, we use the idea of contrast learning to learn modal features.
- **ACMR** [14]. It is based on an adversarial learning approach and triplet constraints to learn representations which are both discriminative and modality-invariant for cross-modal retrieval.
- **Corr-AE and Corr-Cross-AE** [3]. It learns the common space features via incorporating autoencoder cost with correlation cost into a single process. Corr-AE is extended to Corr-Cross-AE by replacing the basic autoencoder with a cross-modal autoencoder.
- **DBRLM** [8]. The model is characterized by deep and bidirectional representation learning. The learning objective is to increase the similarity of matched pairs and reduce the similarity of unmatched pairs, which is achieved by using the framework of maximum likelihood.

Ablation Models

- **TrajCross/nogan**. We removed the adversarial learning module so that we can verify its impact on the model.
- **TrajCross/noxy**. We removed the location information of the image in the model to illustrate the role of the information.

TABLE I
The results of different methods(%)

Method	trajectory2image			image2trajectory		
	top@1	top@5	top@10	top@1	top@5	top@10
CCA	0.509	1.32	2.12	0.04	1.296	2.509
MOCO	4.29	14.29	22.78	2.82	13.87	21.94
ACMR	0.42	0.21	1.32	0.38	0.32	1.44
Corr-AE	0.04	0.31	0.42	0.08	0.25	0.34
Corr-Cross-AE	0.04	0.21	0.46	0.04	0.21	0.46
DBRLM	0.04	0.21	0.41	0.12	0.71	1.43
TrajCross	69.8	90.4	94.3	70.2	91.1	94.5

A. *RESULTS AND ANALYSIS*

1) Performance Comparison: We compare TrajCross with all the baseline methods. Table I shows the experimental

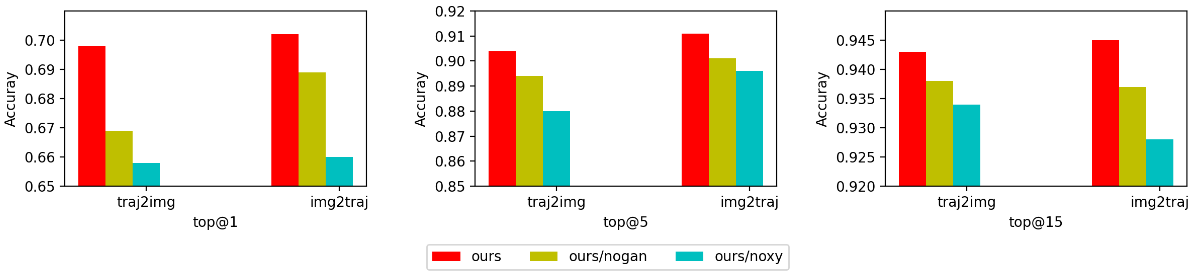

Fig. 3. The ablation experimental results of cross-model retrieval. Overall, the model proposed in this paper has the highest accuracy compared with other variants in terms of top@1, top@5, top@15.

results. According to the results, we have two key observations. First, we can observe that the experimental results of our model are significantly better than other methods, which proves the effectiveness of our method in cross-modal retrieval scenarios of trajectories. Traditional methods cannot play a role in the multi-modal retrieval of trajectories because of their limited ability to discover data relationships. The existing multi-modal retrieval methods based on deep learning cannot achieve good results in the multi-modal retrieval scene of the trajectory, mainly because the shape and position information of the trajectory cannot be well represented. Second, among these benchmark methods, the Moco method is better than other methods, mainly because the method is based on images to achieve cross-modal retrieval and can capture the shape information of the image. This shows that it is useful to extract trajectory information from images, which provides another way to process trajectories.

TABLE II
The experimental results of various variants(%)

Method	trajectory2image			image2trajectory		
	top@1	top@5	top@10	top@1	top@5	top@10
TrajCross/OnlyTraj	3.45	13.3	20.8	2.4	11.1	20.3
TrajCross/TrajCNN	0.04	0.21	0.42	0.08	0.25	0.42

2) Ablation study: We perform ablation experiments to analyze TrajCross by comparing it with two ablations. The results are shown in Figure 3.

According to Figure 3, overall, the performance improves compared to TrajCross/nogan that does not use GAN when training the model In terms of top@1, top@5 and top@15. Taking the trajectory retrieval image as an example, the accuracy is increased by 2.9%, 1.0%, 0.5% respectively. This means that adversarial learning helps to eliminate the discrepancy between modalities.

When comparing TrajCross/noxy with TrajCross/nogan and TrajCross, we can see that the accuracy of TrajCross/noxy is obviously lower. Specifically, taking the trajectory retrieval image as an example, the accuracy are reduced by 4.0%, 2.4%, 0.9% respectively compared with TrajCross in terms of top@1, top@5 and top@15. This shows that in the cross-modal retrieval of trajectories, the model must consider the geographic information covered by the image, otherwise the obtained results will have errors.

In order to be able to further confirm which factor of the image and the trajectory is the key factor for the successful cross-modal retrieval of the trajectory, we conducted two following experiments respectively:

- **TrajCross/onlyTraj**. We only use the trajectory sequence and do not use the image corresponding to the trajectory as the input of the model to obtain the results of cross-modal retrieval.
- **TrajCross/TrajCNN**. In order to further verify the effectiveness of the image corresponding to the trajectory in cross-modal retrieval, we replace this module with the original trajectory sequence to extract the shape information directly. We use CNN to extract the shape of the trajectory sequence.

The results are shown in Table II. Through the experimental results of TrajCross/onlyTraj, we can see that it is not accurate to achieve cross-modal retrieval based on the information provided by the trajectory sequence alone. Meanwhile, we can clearly see that the accuracy of TrajCross/TrajCNN is significantly lower than that of TrajCross, which means that images are the key to successful cross-modal retrieval. The main reason for this situation is that in the process of cross-modal retrieval for trajectory, we can extract more features to assist cross-modal retrieval through images, but it is difficult to learn useful features directly through trajectory sequence. This provides a new idea for the processing of trajectory sequences for subsequent research work.

V. Conclusion and Future Work

In this paper, we define a new research problem *i.e.*, trajectory cross-modal retrieval. The challenge is how to effectively embed the data of each modality, and at the same time, reduce or eliminate the cross-domain discrepancy caused by the inconsistent data distribution. We adopt the method of multi-level feature extraction to obtain the shape and position feature representation of each modality. In order to eliminate the gap of multi-modal, we apply the method of adversarial learning. Comprehensive experimental results and extensive analysis have proved the effectiveness of our method. Compared with the most advanced methods, it can provide excellent performance on cross-modal retrieval.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China under Grant No.: 62077044, 61702470, 62002343.

References

[1] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 961–971 (2016)

[2] Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International conference on machine learning. pp. 1247–1255. PMLR (2013)

[3] Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 7–16 (2014)

[4] Giuliari, F., Hasan, I., Cristani, M., Galasso, F.: Transformer networks for trajectory forecasting. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 10335–10342. IEEE (2021)

[5] Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (2010)

[6] Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural computation **16**(12), 2639–2664 (2004)

[7] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)

[8] He, Y., Xiang, S., Kang, C., Wang, J., Pan, C.: Cross-modal retrieval via deep and bidirectional representation learning. IEEE Transactions on Multimedia **18**(7), 1363–1377 (2016)

[9] Jia, Y., Salzmann, M., Darrell, T.: Learning cross-modality similarity for multinomial data. In: 2011 International Conference on Computer Vision. pp. 2407–2414. IEEE (2011)

[10] Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., Wu, Y.: Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410 (2016)

[11] Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Computer graphics forum. vol. 26, pp. 655–664. Wiley Online Library (2007)

[12] Pereira, J.C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE transactions on pattern analysis and machine intelligence **36**(3), 521–535 (2013)

[13] Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: International conference on machine learning workshop. vol. 79, p. 3 (2012)

[14] Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 154–162 (2017)

[15] Yao, D., Cong, G., Zhang, C., Bi, J.: Computing trajectory similarity in linear time: A generic seed-guided neural metric learning approach. In: 2019 IEEE 35th international conference on data engineering (ICDE). pp. 1358–1369. IEEE (2019)

[16] Yao, D., Zhang, C., Zhu, Z., Huang, J., Bi, J.: Trajectory clustering via deep representation learning. In: 2017 international joint conference on neural networks (IJCNN). pp. 3880–3887. IEEE (2017)

[17] Zhu, B., Ngo, C.W., Chen, J., Hao, Y.: R2gan: Cross-modal recipe retrieval with generative adversarial network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11477–11486 (2019)