Multi-modal Graph Contrastive Learning for Micro-video Recommendation

Zixuan Yi
z.yi.1@research.gla.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

Xi Wang
x.wang.6@research.gla.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

Iadh Ounis
iadh.ounis@glasgow.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

Craig Macdonald
craig.macdonald@glasgow.ac.uk
University of Glasgow
Glasgow, Scotland, United Kingdom

ABSTRACT

Recently micro-videos have become more popular in social media platforms such as TikTok and Instagram. Engagements in these platforms are facilitated by multi-modal recommendation systems. Indeed, such multimedia content can involve diverse modalities, often represented as visual, acoustic, and textual features to the recommender model. Existing works in micro-video recommendation tend to unify the multi-modal channels, thereby treating each modality with equal importance. However, we argue that these approaches are not sufficient to encode item representations with multiple modalities, since the used methods cannot fully disentangle the users' tastes on different modalities. To tackle this problem, we propose a novel learning method named Multi-Modal Graph Contrastive Learning (MMGCL), which aims to explicitly enhance multi-modal representation learning in a self-supervised learning manner. In particular, we devise two augmentation techniques to generate the multiple views of a user/item: *modality edge dropout* and *modality masking*. Furthermore, we introduce a novel negative sampling technique that allows to learn the correlation between modalities and ensures the effective contribution of each modality. Extensive experiments conducted on two micro-video datasets demonstrate the superiority of our proposed MMGCL method over existing state-of-the-art approaches in terms of both recommendation performance and training convergence speed.

ACM Reference Format:
Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal Graph Contrastive Learning for Micro-video Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3532027

1 INTRODUCTION

Micro-videos, a very ubiquitous medium, allow users to share over social media platforms (e.g. TikTok, Instagram). Abundant multimedia content is uploaded to the platforms and interacted with users. This content involves diverse modalities of visual, acoustic, and textual features. As such, a myriad of recommendation algorithms [1, 4, 7, 9] have been proposed to incorporate multi-modal information into the collaborative filtering (CF) scheme, especially through the development of Graph-based CF methods [20, 24]. Specifically, some previous studies [15, 20] incorporated multi-modal information to generate node representation in a uni-graph. Alternatively, Wei et al. [18] leveraged the learned user preferences from each modality graph respectively, but without disentangling the users' tastes on different modalities. Consequently, the aforementioned works [15, 18, 20] – which either unified or homogenized multi-modal channels – treated each modality with equal importance. This might lead to suboptimal representations for example by overemphasizing a particularly modality in the learned representations. Inspired by recent studies [19, 22, 25], which have shown the superior ability of Self-Supervised Learning (SSL) to construct supervised signals from correlation within raw data, this paper investigates the possibility of leveraging SSL to explore the correlations among modalities and alleviate the equal importance problem in micro-video recommendations. On the other hand, contrastive learning has recently become a dominant component in SSL. A typical way [19] to apply contrastive learning to recommendation on graphs is by generating multiple views by perturbing the user-item bipartite graphs. Then the views are contrasted by maximizing the agreement between different views of the same node (i.e. positive pairs) and increasing the disagreement compared to other nodes (i.e. negative pairs). However, existing SSL approaches based on graphs (e.g. SGL [19]) cannot be directly applied to micro-video recommendations because they fail to generate informative views for nodes from multiple modalities. In addition, they cannot provide an estimation of the users' tastes on different modalities as auxiliary signals during training.

To address the above problem, we propose a novel learning method called Multi-Modal Graph Contrastive Learning (MMGCL) to explicitly enhance the multi-modal representation learning in micro-video recommendations. For this purpose, MMGCL leverages positive pairs <anchor sample, positive sample> of nodes by devising two augmentation techniques: *modality edge dropout* and *modality masking*. The first removes edges from graphs of different modalities and the second technique selectively masks one particular modality of user/item features. Furthermore, MMGCL generates challenging negative samples by perturbing one particular modality of the positive sample. The joint contribution of the above techniques encourages the encoder to learn the correlation from different modalities, ensuring the effective contribution of each modality.

To summarise, our contributions are threefold: (1) We propose a new self-supervised graph learning method for the micro-video recommendation, which combines the traditional pairwise ranking

SIGIR '22, July 11–15, 2022, Madrid, Spain

task objectives with contrastive learning objectives. (2) We devise two multi-modal-specific augmentation techniques to construct the multi-views of a node. To ensure the effective contribution of each modality, we propose an effective negative sampling strategy by perturbing one of the modalities. (3) We conduct extensive experiments on two public datasets and demonstrate that MMGCL outperforms multiple strong baselines while especially enhancing the training convergence speed.

2 RELATED WORK

Graph-based CF recommender. Graph-based recommenders [8, 10, 17] principally exploit high-order connectivity in the user-item graph by propagating information from local neighbours and integrate the collaborative signals into the user/item representation. However, existing approaches (e.g. LightGCN [8]) only propagate homogeneous features from a single data source, which does not allow to leverage the correlation between different modalities. Hence, we encode user/item features from modal-specific graphs with the aid of LightGCN and feed them into a contrastive learning framework to improve micro-video recommendation performance.

Multi-modal recommendation. In multi-modal recommendation, the content information of items is incorporated into a CF-based schema to yield better item representations [9, 20]. Some approaches [1, 4, 7] leverage uni-modal features to enrich item representations for various recommendations. Moreover, MMGCN [18] leverage visual, acoustic and textual features in parallel to model the users' preferences. Furthermore, prior works [3, 5, 18] have noted the importance of leveraging the modalities' information to enhance the recommendation results but they failed to leverage the correlation between multiple modalities. In contrast, our work fully explores the correlation between modalities to enhance the multi-modal representation learning via SSL.

Self-supervised learning for recommendation. Recently, a number of self-supervised learning recommenders [19, 22, 23, 25, 26] have applied various data augmentations to improve recommendation performance. A recent SSL method, SGL [19], proposed use of node dropout, edge dropout and random walk augmentations. However, the above augmentations on uni-graphs cannot generate informative views of a node that fully inherit the rich information from multiple modalities. Moreover, none of the aforementioned recommenders explored an effective negative sampling strategy in contrastive learning. Hence, our work focuses on investigating modal-specific graph augmentations to address this gap, by generating informative views of nodes from multiple modalities and a negative sampling strategy to enhance effectiveness in contrastive learning.

3 METHODOLOGY

3.1 Problem Definition and Notations

Following [18], we devise a bipartite graph for each modality, where nodes represent users/micro-videos and edges indicate interactions between users and micro-videos. More formally, we use $m \in \mathcal{M} = \{v, a, t\}$ as the modality indicator, where v, a, and t represent the visual, acoustic, and textual modalities, respectively. Given the multi-modal interaction graphs $\mathcal{G}_v$, $\mathcal{G}_a$ and $\mathcal{G}_t$, we aim to estimate user preferences through a multi-modal encoder f that can recommend the top-k micro-videos for a target user u.

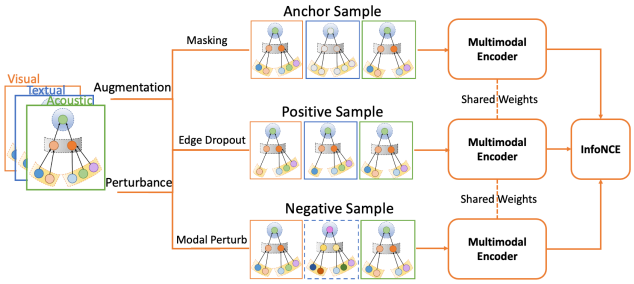

Figure 1: Overview of our proposed MMGCL method

3.2 Multi-view Graph Augmentation

Inspired by [6, 16], who applied multi-view representation learning for contrastive learning, we devise two augmentation operators on the multi-modal graphs $\mathcal{G}_v$, $\mathcal{G}_a$ and $\mathcal{G}_t$, namely *modality edge dropout* and *modality masking*, to create a multi-view representation $V(\mathcal{G})$ of each node. Figure 1 provides an overview of our proposed framework.

Modality Masking: As illustrated in Figure 1, we apply a masking pattern on a particular modality of user/item features as follows:

$$V_1(\mathcal{G}) = \begin{cases} (\mathcal{V}_v, \mathcal{E}_v) \parallel (\mathcal{V}_a, \mathcal{E}_a) \parallel (M_1 \odot \mathcal{V}_t, \mathcal{E}_t) & \text{with } p_t \\ (\mathcal{V}_v, \mathcal{E}_v) \parallel (M_1 \odot \mathcal{V}_a, \mathcal{E}_a) \parallel (\mathcal{V}_t, \mathcal{E}_t) & \text{with } p_a \\ (M_1 \odot \mathcal{V}_v, \mathcal{E}_v) \parallel (\mathcal{V}_a, \mathcal{E}_a) \parallel (\mathcal{V}_t, \mathcal{E}_t) & \text{with } p_v \end{cases} \quad (1)$$

where $\parallel$ represents the concatenation operator, p_v, p_a, p_t are the individual probabilities to control which of the three modalities will be masked and $p_v + p_a + p_t = 1$; M_1 is the masking vector on the node set $\mathcal{V}$. We implement this masking operator by replacing a particular modality of user/item features with a randomly initialized embedding in the input layer. The masking step can be interpreted as a special case of a 100% dropout rate. As such, this augmentation is expected to increase the contribution of each modality with the consistent absence of the masked modalities during training.

Modality Edge Dropout: This randomly removes edges in each modality graph with a dropout ratio ρ. The resulting view is represented as:

$$V_2(\mathcal{G}) = (\mathcal{V}_v, M_2 \odot \mathcal{E}_v) \parallel (\mathcal{V}_a, M_2 \odot \mathcal{E}_a) \parallel (\mathcal{V}_t, M_2 \odot \mathcal{E}_t) \quad (2)$$

where $\mathcal{V}$ is the node set and M_2 is the masking vector on edge set $\mathcal{E}$. This derived view is created by a set of sub-graphs from the original multi-modal graphs and can still preserve the users' main intentions on different modalities. As such, this augmentation is expected to capture the useful patterns of a node on each uni-modal graph and further endows the representations by concatenation.

3.3 Challenging Negative Samples

Hard negative mining has been effectively applied in multi-modal fusion scenarios where particular modalities tend to dominate the learned representations [11, 12, 21]. Similarly, we leverage a perturbation strategy to generate negative samples in contrastive learning. Given a collection of samples $\{s_1^i, s_2^i\}_{i=1}^{\mathcal{N}}$, we contrast the positive pair $x = \{s_1^i, s_2^i\}$ and the negative pair $y = \left\{s_1^i, s_2^j\right\}$. For example, given an anchor sample s_1^1 that consists of three modalities $(c_{1,i}^v, c_{1,i}^a, c_{1,i}^t)$ and its positive sample s_2^1 represented as

$(c^v_{2,i}, c^a_{2,i}, c^t_{2,i})$, we propose a perturbed negative sample s^j_2 represented as $(c^v_{2,j}, c^a_{2,d(j)}, c^t_{2,j},)$, where $d(\cdot)$ is a perturbing function producing a random index from the sample set. As a result, the multi-modal encoder has to discriminate between the positive sample and the negative sample with only one modality difference. Thus with the perturbed negative sample, it becomes especially challenging for a network to tell whether the perturbed modality $c^m_{2,d(j)}$ is in correspondence with the rest of modalities or not. Therefore the challenging negative samples encourage learning the correlation of different modalities with the perturbing function.

3.4 Contrastive Learning

Having obtained a positive pair from two randomly augmented views and a negative pair with a positive sample and challenging negatives, we follow SimCLR [2] and adopt the contrastive loss, InfoNCE [13], to maximize the agreement of positive pairs and minimize that of the negative pairs:

$$\mathcal{L}^{user}_{ssl} = -\mathbb{E}_{\{s^1_1, s^1_2, \ldots, s^{k+1}_2\}} \left[\log \frac{h(\{s^1_1, s^1_2\})}{\sum_{j=1}^{k+1} (h(\{s^1_1, s^j_2\})} \right], \tag{3}$$

where k is the number of negative sample s^j_2 for a given anchor sample s^1_1. We compute the similarities of the positive/negative pairs as scores and adjust their dynamic range by a hyper-parameter τ:

$$h(\{s^1_1, s^1_2\}) = \exp\left(\frac{f(V_1(\mathcal{G})) \cdot f(V_2(\mathcal{G}))}{\|f(V_1(\mathcal{G}))\| \cdot \|f(V_2(\mathcal{G}))\|} \cdot \frac{1}{\tau}\right) \tag{4}$$

where f is the multi-modal encoder to extract compact latent representations of $V_1(\mathcal{G})$ and $V_2(\mathcal{G})$ We simply fix one view and enumerate positives and negatives from the other view. The loss function in Equation (3) treats a view $V_1(\mathcal{G})$ as an anchor and enumerates over $V_2(\mathcal{G})$. Symmetrically, we can obtain the loss by anchoring at $V_2(\mathcal{G})$ and add them up as our two-views loss. Analogously, we obtain the contrastive loss of the item side $\mathcal{L}^{item}_{ssl}$. Combining these two losses, we obtain an objective function for the self-supervised task as: $\mathcal{L}_{ssl} = \mathcal{L}^{user}_{ssl} + \mathcal{L}^{item}_{ssl}$.

3.5 Multi-task Training

To improve the recommendations with contrastive learning, we adopt a multi-task training strategy to jointly optimize the pairwise ranking task objectives and the contrastive learning objective $\mathcal{L}_{ssl}$:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ssl} + \sum_{(u,i,j) \in D_s} \ln \sigma(y_{ui} - \boldsymbol{e_u}^{\mathrm{T}} \boldsymbol{e_i}) + \lambda_2 \|\Theta\|^2_2 \tag{5}$$

where the second term is the adopted Bayesian Personalized Ranking (BPR) loss [14], $\boldsymbol{e_u}$ is the user embedding, $\boldsymbol{e_i}$ denotes the positive item embedding and y_{ui} is the ground truth value, $D_s = \{(u,i,j)|(u,i) \in R^+, (u,j) \in R^-\}$ is the set of the training data, R^+ indicates the observed interactions and R^- indicates the unobserved interactions, $\sigma(\cdot)$ is the sigmoid function, Θ is the set of model parameters in the BPR loss since L_{ssl} introduces no additional parameters, while λ_1 and λ_2 are hyperparameters to control the strengths of SSL and L_2 regularization, respectively. It is worthwhile to note that the challenging negatives in contrastive learning is distinct with negatives $\boldsymbol{e_j}$ in the BPR loss.

Table 1: Statistics of the TikTok and MovieLens datasets.

	TikTok	MovieLens
Users	48,524	12,674
Items	84,236	4,214
Interactions	4,751,504	1,013,573
Interaction density(%)	0.0012	0.0084
Visual Dimension	128	128
Acoustic Dimension	128	128
Textual Dimension	128	100

Table 2: Summary of approaches across different aspects.

Method	NGCF	LightGCN	MMGCN	SGL	MMGCL
Graph-CF	✓	✓	×	×	✓
Multi-modal	×	×	✓	×	✓
SSL-based	×	×	×	✓	✓

4 EXPERIMENTS

To demonstrate the effectiveness of MMGCL and illustrate the reasons for this effectiveness, we conduct experiments to answer the following research questions:

RQ1: (a) How does MMGCL perform micro-video top-K recommendation compared with baselines? (b) How do different augmentations and the negative sampling strategy impact the performance?

RQ2: Can we leverage MMGCL to achieve faster convergence speed compared to the baselines?

4.1 Experimental Setting

4.1.1 Datasets. To evaluate our MMGCL method, we conduct experiments on two public micro-video datasets: *TikTok*[1] and *MovieLens*[2]. The statistics of the datasets are listed in Table 1. For each dataset, we remove the users with less than ten interactions and preserve items with more than ten interactions. Moreover, we proceed with a dimension reduction operation [3] on the visual features from a 1024 dimension vector to 128 dimensions in the MovieLens dataset to reduce the redundancies of the embeddings.

4.1.2 Experimental Setup. To evaluate the effectiveness of our model, we compare MMGCL with the following state-of-the-art baselines which were discussed in Section 2: **NGCF** [17], **LightGCN** [8], **MMGCN** [18] and **SGL** [19]. Table 2 compares MMGCL to the baselines across different aspects.

4.1.3 Evaluation Protocol and Hyper-parameter Settings. We randomly split a given dataset into training, validation, and testing sets with 8:1:1 ratio. We adopt two widely used evaluation metrics: Recall@K and NDCG@K to evaluate the performance of top-K recommendations. We set K = 10 and report the averaged performance achieved for all users in the testing set. The negative items of each user are defined as those having no interactions with the user. We adopt the Xavier initialization to initialize all the model parameters and use Adam optimizer for model optimization with a batch size of 1024. We apply early-stopping during training, terminating the training when the validation loss does not decrease for 50 epochs. Moreover, we tune the hyper-parameters on the validation set. The learning rate is selected from $\{10^{-2}, 10^{-3}, 10^{-4}\}$.

[1] http://ai-lab-challenge.bytedance.com/tce/vc/

[2] https://grouplens.org/datasets/movielens/

Table 3: Effectiveness of MMGCL and the baselines. Improvements over the baselines are statistically significant with a p−value < 0.01 using the student's t-test.

Dataset	TikTok		MovieLens	
Methods	Recall@10	NDCG@10	Recall@10	NDCG@10
NGCF	0.1783	0.3861	0.4376	0.3162
LightGCN	0.1896	0.4323	0.4695	0.3381
MMGCN	0.1935	0.4315	0.4684	0.3359
SGL	0.1951	0.4357	0.4702	0.3510
MMGCL	**0.2067**	**0.4681**	**0.4943**	**0.3713**
%Improve.	5.95%	7.44%	5.13%	5.77%
p-value	$4.34e-10$	$5.83e-10$	$1.97e-7$	$3.46e-7$

For those hyper-parameters unique to MMGCL, we tune λ_1, τ and ρ within the ranges of $\{0.1, 0.2, 0.5, 1.0\}$, $\{0, 0.1, 0.2, ..., 1.0\}$ and $\{0, 0.1, 0.2, ..., 0.9\}$, respectively. Moreover, we also tune p_v, p_a, p_t in the same range of $\{0, 0.1, 0.2, ..., 1.0\}$.

4.2 MMGCL Effectiveness Evaluation (RQ1)

4.2.1 Performance Comparison with Baselines. To evaluate our proposed method, we report the empirical results of all baselines and the improvements in respect of each component of MMGCL, which are calculated between our proposed method and the strongest baselines highlighted with underline, in Table 3. By comparing with NGCF and LightGCN, we validate the argument in the literature that graph-based CF methods cannot disentangle the users' tastes on different modalities. Next, MMGCL outperforms MMGCN by a large margin, which demonstrates the rationality and effectiveness of incorporating self-supervised learning in the micro-video recommendation method. One possible reason for this phenomenon is that certain modalities of MMGCN dominate in the learned representations and the rest of modalities are ignored. The fact that MMGCL outperforms SGL reveals the effectiveness of the multimodal specific augmentations and the negative sampling strategy. Finally, according to the results, MMGCL consistently yields the best performance across the table. Hence, we answer RQ1(a) as follows: The significant improvements from MMGCL is attributed to the mutual supplement of efficient representation learning from informative views and learning the correlation of different modalities with challenging negatives in an SSL schema.

4.2.2 Effect of Multi-view Augmentation and Challenging Negatives. To explore the effects of different augmentations and the negative sampling strategy, we compare the results in Figure 2 to conclude on the effectiveness of multi-view augmentation and the proposed negative sampling method. We use ED/MM/CN as the abbreviations of Modality Edge Dropout, Modality Masking and Challenging Negatives, respectively. As expected, the augmentation method with multi-view outperforms those with single views in MMGCL. It demonstrates the successful exploitation of the fundamental supervisory signal, namely the co-occurrence of multiple views of the users' preferences. For the comparison between single-view augmentations, the performance of $MMGCL_{MM}$ is on a par with $MMGCL_{ED}$ on TikTok but not competitive with $MMGCL_{ED}$ on Movielens. One possible reason is that TikTok has more effective

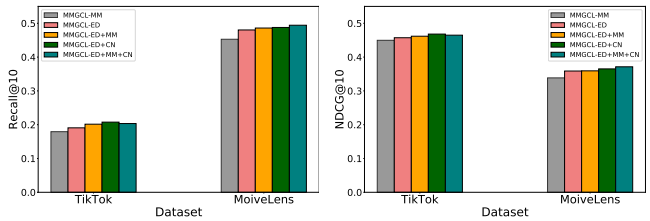

Figure 2: Performances in terms of Recall@10 and NDCG@10 on MovieLens and TikTok.

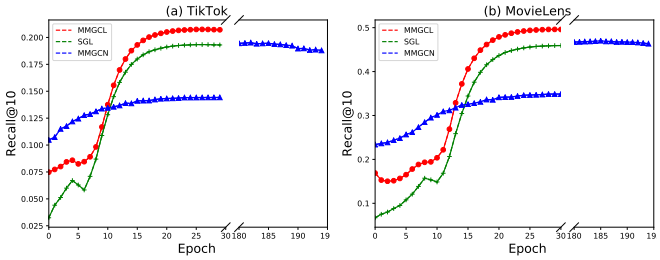

Figure 3: Training curves of MMGCL, SGL and MMGCN on MovieLens and TikTok.

embeddings than MovieLens and $MMGCL_{MM}$ masks the most effective content from multiple modalities. This is further verified in that MMGCL achieves more improvements on TikTok than Movielens with the joint contribution of the dropout and modality masking in Table 3. Moreover, $MMGCL_{ED}$ outperforms $MMGCL_{MM}$ in most cases, which indicates that perturbing the graph structure can capture more useful inherent patterns on the user's potential interests. Next, the corresponding results show that the improvements come from the proposed challenging negatives $MMGCL_{ED+MM+CN}$ by comparing to $MMGCL_{ED+MM}$, which only applies the augmentation views. This reveals that our method provides more factual negative samples by the modal-specific perturbing strategy. Hence, we answer RQ1(b) as follows: MMGCL successfully leverages the multi-view augmentation to learn effective representations and further enhances performance by facilitating the learning of correlations among modalities with the challenging negatives.

4.3 Training Efficiency (RQ2)

Previous work [19] has shown the superiority of self-supervised learning in training efficiency. Thus, we further study the training efficiency on the implementations of multi-view augmentation and the challenging negative samples. Figure 3 shows the training curves of MMGCL, SGL and MMGCN on the TikTok and MovieLens datasets. We observe that MMGCL is much faster to converge than MMGCN on both datasets. In particular, MMGCL arrives at the best performance after 20 epochs, while MMGCN takes more than 180 epochs in these two datasets, respectively. This suggests that our proposed MMGCL method can greatly reduce the training time compared to MMGCN, meanwhile achieving significant improvements and outperforming SGL through all epochs in the figure. Hence, we answer RQ2 as follows: MMGCL successfully leverages

the representation learning by multi-view augmentation and provides large gradients during training by contrasting challenging negative samples.

5 CONCLUSIONS

In this work, we tackled the equal importance problem in micro-video recommendations with a self-supervised learning paradigm and explored the potential of SSL in enhancing multi-modal representation learning in top-k micro-video recommendation. From the perspective of multi-modal user-item graphs, we devised two data augmentations from different aspects to generate informative views in the auxiliary contrastive task. From the perspective of negative sampling, we proposed a perturbing strategy to generate challenging negative samples to fully explore the correlations among modalities and to ensure the effective contribution of each modality in the learned representations. Furthermore, we conducted extensive experiments on two benchmark datasets to justify the advantages of our proposed MMGCL method in micro-video recommendation in terms of improving the training performance and convergence speed.

REFERENCES

[1] Tao Chen, Xiangnan He, and Min-Yen Kan. 2016. Context-aware image tweet modelling and recommendation. In *Proceedings of the 24th ACM international conference on Multimedia*. 1018–1027.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 1597–1607.

[3] Xiaoyu Du, Xiang Wang, Xiangnan He, Zechao Li, Jinhui Tang, and Tat-Seng Chua. 2020. How to learn item representation for cold-start multimedia recommendation?. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3469–3477.

[4] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2017. A unified personalized video recommendation via dynamic recurrent neural networks. In *Proceedings of the 25th ACM international conference on Multimedia*. 127–135.

[5] Xuri Ge, Fuhai Chen, Joemon M Jose, Zhilong Ji, Zhongqin Wu, and Xiao Liu. 2021. Structured multi-modal feature embedding and alignment for image-sentence retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5185–5193.

[6] Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *Proceedings of 37th International Conference on Machine Learning*. PMLR, 4116–4126.

[7] Ruining He and Julian McAuley. 2016. VBPR: Visual bayesian personalized ranking from implicit feedback. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, Vol. 30.

[8] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[9] Shuhui Jiang, Xueming Qian, Jialie Shen, Yun Fu, and Tao Mei. 2015. Author topic model-based collaborative filtering for personalized POI recommendations. *IEEE Transactions on Multimedia* 17, 6 (2015), 907–918.

[10] Siwei Liu, Iadh Ounis, Craig Macdonald, and Zaiqiao Meng. 2020. A heterogeneous graph neural model for cold-start recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2029–2032.

[11] Yunze Liu, Li Yi, Shanghang Zhang, Qingnan Fan, Thomas Funkhouser, and Hao Dong. 2020. P4Contrast: Contrastive learning with pairs of point-pixel pairs for RGB-D scene understanding. *arXiv preprint arXiv:2012.13089* (2020).

[12] Zijun Long and Richard McCreadie. 2022. Is multi-modal data key for crisis content categorization on social media?. In *Proceedings of 19th International Conference on Information Systems for Crisis Response and Management*.

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[14] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*. 452–461.

[15] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1405–1414.

[16] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Proceedings of the 16th European Conference on Computer Vision*. Springer, 776–794.

[17] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 165–174.

[18] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.

[19] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 726–735.

[20] Jiaxin Wu, Sheng-Hua Zhong, and Yan Liu. 2019. Mvsgcn: A novel graph convolutional network for multi-video summarization. In *Proceedings of the 27th ACM International Conference on Multimedia*. 827–835.

[21] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Proceedings of the 16th European Conference on Computer Vision*. Springer, 574–591.

[22] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Bolin Ding, and Bin Cui. 2020. Contrastive learning for sequential recommendation. *arXiv preprint arXiv:2010.14395* (2020).

[23] Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, et al. 2020. Self-supervised learning for deep models in recommendations. *arXiv e-prints* (2020), arXiv–2007.

[24] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. 2018. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 974–983.

[25] Junliang Yu, Hongzhi Yin, Jundong Li, Qinyong Wang, Nguyen Quoc Viet Hung, and Xiangliang Zhang. 2021. Self-supervised multi-channel hypergraph convolutional network for social recommendation. In *Proceedings of the 30th Web Conference 2021*. 413–424.

[26] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 1893–1902.