Self-Supervised Learning for Multimedia Recommendation

Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, *Member, IEEE*, Lifang Yang, Xianglin Huang, and Tat-Seng Chua

***Abstract*—Learning representations for multimedia content is critical for multimedia recommendation. Current representation learning methods roughly fall into two groups: (1) using the historical interactions to create ID embeddings of users and items, and (2) treating multi-modal data as the side information of items to enrich their ID embeddings. Each user-item interaction offers the supervisory signal to optimize the representation learning by the traditional supervised learning paradigm. Due to the overlook of the multi-modal patterns ($e.g.$, co-occurrence of visual, acoustic, textual features in micro-videos a user saw before, and her behavioral features) hidden in the data, these methods are insufficient to create powerful representations and obtain satisfactory recommendation accuracy. To capture multi-modal patterns in the data itself, we go beyond the supervised learning paradigm, and incorporate the idea of self-supervised learning (SSL) into multimedia recommendation. Specifically, SSL consists of two components: (1) data augmentation upon multi-modal contents, where we design three operators — feature dropout (FD), feature masking (FM), feature fine and coarse spaces (FAC) — to generate multiple views of individual items; and (2) contrastive learning, which differentiates the views of an item from the others' to distill additional supervisory signals. Clearly, SSL enables us to explore and exhibit the underlying relations among modalities, thereby resulting in powerful representations. We denote the generic framework by *Self-supervised Learning-guided Multimedia Recommendation* (SLMRec). Extensive experiments are performed on three real-world datasets, showing that SLMRec achieves significant improvements over several state-of-the-art baselines like LightGCN [1], MMGCN [2]. Further analysis shows how SSL affects recommendation performance.**

***Index Terms*—Multimedia recommendation, self-supervised learning, graph neural network, micro-videos.**

I. Introduction

Multimedia recommendation has been widely employed in a wide range of real-world applications, such as E-commerce, micro-video sharing platforms. The rich multimedia contents ($e.g.$, visual, acoustic, textual features of micro-videos) can supplement the historical user-item interactions ($e.g.$, views, purchases, clicks). These multi-modal data not only reflect content relatedness among items ($e.g.$, having similar background music) but also indicate users' finer-grained preference at the granularity of modalities. Furthermore, there indeed exist some underlying relationships among these modalities, such as similar and consistent semantics reflecting in visual and acoustic modalities. Hence, individually treating them will ignore their entangled relationships.

Current multimedia recommender models mostly follow a supervised learning paradigm — transfer user behaviors with multi-modal data into generic feature vectors ($i.e.$, representations), and feed them into a supervised learning model guided by the supervisory signal ($i.e.$, historical interactions). Clearly, learning quality representations is of great importance to infer user preference accurately. Current representation learning methods roughly fall into two research lines: (1) using historical interactions solely to create ID embeddings, so as to encode collaborative signal among users; and (2) treating multi-modal data as the side information of item features to enrich ID embeddings, with the aim to reflect content similarities. The first line, also well-known as collaborative filtering, typically utilizes neural networks as the encoder to generate ID embeddings. The encoder evolves from early latent factors upon individual IDs [3], [4], MLPs upon personal historical [5], [6], to recent graph neural networks upon the holistic interaction graph [1], [7]. Founded upon the first line, the second line focuses on integrating the multi-modal features with ID embeddings. Typically, the multi-modal features are extracted by pre-trained neural networks, such as feeding keyframes of micro-videos into ResNet [8] to obtain visual features. Early integrating strategies like VBPR [9] focus on individual items and aggregate multi-modal features and ID embedding of an item into its representation via concatenation or summation operations. Some follow-up works like ACF [10] consider personal history, and combine the multi-modal features of historical items adopted by a user as her representation. Recent works like MMGCN [2] generate modality-aware interaction graphs and apply graph neural networks on these graphs holistically to perform information propagation and fusion.

Manuscript received 17 February 2022; revised 29 April 2022 and 2 June 2022; accepted 15 June 2022. Date of publication 30 June 2022; date of current version 30 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFB1406800, in part by the Fundamental Research Funds for the Central Universities under Grants CUC22GZ003, CUC22GZ007, and CUC22GZ035, and in part by GH-fund B under Grants 20220202 and ghfund202202023853. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Wen-Huang Cheng. *(Corresponding author: Xiang Wang.)*

Zhulin Tao, Xiaohao Liu, Yewei Xia, Lifang Yang, and Xianglin Huang are with the State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China (e-mail: taozhulin@gmail.com; liuxiaohao@cuc.edu.cn; yeweixia@cuc.edu.cn; yanglifang@cuc.edu.cn; huangxl@cuc.edu.cn).

Xiang Wang is with the CCCD Key Laboratory of Ministry of Culture and Tourism, University of Science and Technology of China, Hefei, Anhui 230026, China (e-mail: xiangwang@u.nus.edu).

Tat-Seng Chua is with the National University of Singapore, Kent Ridge, Singapore 117543 (e-mail: chuats@comp.nus.edu.sg).

Digital Object Identifier 10.1109/TMM.2022.3187556

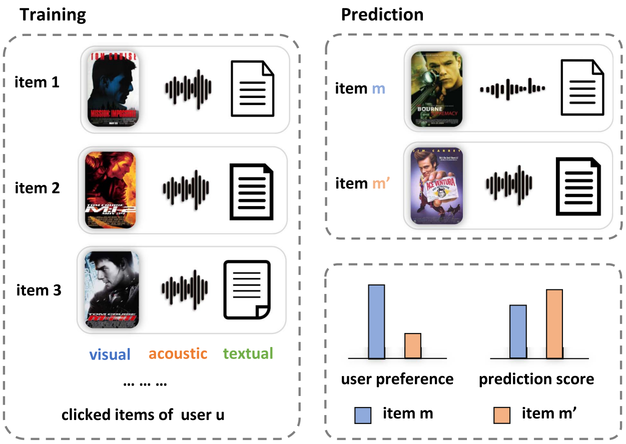


Fig. 1. An example on untouched multi-modal patterns.

However, existing multi-modal recommenders mostly follow two paradigms: early fusion of multi-modal features or late fusion of multi-modal predictions. These fusions focus solely on the observed data, thus suffer from two certain issues: (a) heavily relying on the observed user-item interactions would offer insufficient supervisory signals to guide the representation learning of items; (b) simply fusing the observed multi-modal patterns hardly generalizes to the unseen or noisy multi-modal patterns, which usually happen in the cold-start scenarios with newly-coming items. As shown in Fig. 1, co-occurring modalities ($e.g.$, video, audio, and dialogue) jointly profile a movie, as well as its audience users. Given a mono-modal content of a movie ($e.g.$, observed video), one can possibly predict or imagine the rest of the contents ($e.g.$, relevant but unobserved music). Specifically, user u clicked a series of items (movies), most of which can be intuitively categorized into spy fiction like *Mission: Impossible Series*. These movies have a similar kind of acoustic soundtrack but with different scenes and different dialogues. Thus, the model will learn from the statistic of whether exists similar acoustic features during the training rather than uncovering multi-modal patterns in these spy fiction films for the basis of preference judgment. However, given item m and m', where m is a classic spy fiction type film and m' is a comedy with *Mission: Impossible* music, the trained model would recommend m' that is contrary to the real preferences of user u. Hence, overlooking such patterns makes these methods fall short in revealing hidden information and results in suboptimal representations.

In this work, we go beyond the supervised learning paradigm and incorporate the idea of self-supervised learning to explore and exploit the multi-modal patterns. After the data augmentations customized for multimedia contents, we obtain more supervisory signals, which reflect the relationships among multiple modalities, and further endow the recommender with better robustness and generalization ability. The basic idea of self-supervised learning (SSL) [11]–[13] is to uncover hidden patterns and generate additional supervisions from data itself, without relying on labels. Technically, contrastive SSL consists of two key components: (1) data augmentation, which generates multiple views to describe an individual item; and (2) contrastive learning, which maximizes the agreement between various views of the same item while minimizing the agreement between different items. Hence, it is a natural tool to mine the multi-modal patterns of items and highlight the relations among different modalities. However, to the best of our knowledge, SSL is less explored in multimedia recommendations.

Towards this end, we propose a multi-task learning framework, where the SSL task supplements the supervised learning task in multimedia recommendation. Specifically, we select a graph neural network-based recommender model to serve the primary supervised learning task. Towards the SSL task, we devise three data augmentation operators: feature dropout (FD), feature masking (FM), and feature fine and coarse spaces (FAC). To encourage the model to learn multi-modal patter rather than seeking the shortcuts from a unimodal feature on preference score, the extracted unimodal features should obey the consistency before fusion. If existed single modality emphasis, the assumed multi-modal features represent only the emphasized unimodal features, failing to uncover the deep information in the representation, namely overfitting and noisy sensitivity. We roughly group the proposed methods into modal-agnostic (*i.e.*, FD and FM) and modal-specific (*i.e.*, FAC). The former creates the broken content of features and encourages the visible factors to predict the broken ones, thus maximizing the mutual information. The latter focus more on modalities themselves, constructing fine and coarse spaces to align every two modalities' features to enhance the consistency for better fusion. We denote this framework of self-supervised learning-guided multimedia recommendation by SLMRec. Experiments are conducted on three datasets and are shown that SLMRec exhibits substantial improvements over the state-of-the-art baselines like LightGCN [1], MMGCN [2]. Further analysis shows how SSL affects recommendation performance. In summary, the key contributions are summarized as follows:

- We devise a generic self-supervised framework for multimedia recommendation, SLMRec. It incorporates SSL into graph neural network-based recommender models, so as to uncover latent patterns from items' various modalities and create powerful representations.
- We exploit the multi-modal pattern uncovering from two perspectives of modal-agnostic and modal-specific and accordingly propose three pretext tasks: FD, FM, and FAC.
- We perform extensive experiments on three datasets to verify the rationality and effectiveness of SLMRec. We have already released the code at.[1]

II. Preliminary

In this paper, we represent matrices by bold upper case letters ($e.g.$, $\mathbf{R}$). And we use calligraphic letters to represent sets ($e.g.$, $\mathcal{U}, \mathcal{I}$). Let $\mathcal{O}$ be the total observed interactions($e.g.$, view, browse, or click) between users and items. We regard the historical interaction between a user and an item as an edge on the graph, then we get a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where node set $\mathcal{V} = \mathcal{U} \cup \mathcal{I}$, edge set $\mathcal{E}$=$\mathcal{O}$, $\mathcal{U}$ and $\mathcal{I}$ denote the user and item

[1] [Online]. Available: https://github.com/zltao/SLMRec/

sets respectively. Besides, we use $\mathcal{N}_u$ to represent the item that has interaction with user $u(i.e.,\ \mathcal{N}_u = \{j|(u,j) \in \mathcal{O}\})$. Similarly, $\mathcal{N}_i = \{j|(j,i) \in \mathcal{O}\}$.

Apart from the visual, acoustic, and textual features of the items, we treat ID as a special modality, especially. For simplicity, we use $m \in \mathcal{M} = \{v, a, t, id\}$ as the modality indicator, where v, a, t, and id represent visual, acoustic, textual and ID modalities, respectively. In terms of different modalities, we split the graph $\mathcal{G}$, and the graph that only keeps the features for modality m is formulated as $\mathcal{G}_m$.

A. Abstract Paradigm of GNN

Graph neural network (GNN) techniques are powerful in capturing the higher-order interaction in graphs via message passing among nodes [7]. The overall scheme can be formulated as follows:

$$\mathbf{Z}_m = H(\mathbf{E}_m, \mathcal{G}_m), \tag{1}$$

where $H(\cdot)$ is the GNN function to encode graph structural information into node representation; $\mathbf{Z}_m \in \mathbb{R}^{|\mathcal{V}|\times d}$ is the output representation matrix of all nodes in modality m, while $\mathbf{E}_m \in \mathbb{R}^{|\mathcal{V}|\times d'}$ is the input feature matrix of all nodes in modality m ($\mathbb{R}$ is a common set of real numbers); d and d' denote the dimension of the output representation vectors and input feature vectors, respectively. According to the function, we divide the whole GNN module into aggregation mechanism, update mechanism, and readout mechanism, which we will elaborate on below.

- *Aggregation Mechanism* is responsible for collecting and aggregating the message of neighbors. After l layers, every node catches the information within its l-hop neighbors. The l-th layer's aggregation is formulated as:

$$\mathbf{a}_{u,m}^{(l)} = f_{\text{aggregate}}(\{\mathbf{z}_{i,m}^{(l-1)}|i \in \mathcal{N}_u\}), \tag{2}$$

  where $\mathbf{a}_{u,m}^{(l)}$ denotes the aggregation of vectorized information from node u's neighborhood $\mathcal{N}_u$ after l layers in modality m and $\mathbf{z}_{i,m}^{(l)}$ denotes the representation of node i after l layers in modality m. $\mathbf{z}_{i,m}^{(0)}$ is initialized to ID embedding $\mathbf{e}_i$. Existing aggregation mechanism can be mainly categorized into mean-pooling operation [14] and attention mechanism [15].
- *Update Mechanism* aims to merge the message of the central node and its neighbors [16]. After the l-th layer, the update mechanism combines a node's previous representation with the information obtained from its l-hop neighbors to generate a new representation, which can be formulated as:

$$\mathbf{z}_{u,m}^{(l)} = f_{\text{combine}}(\mathbf{a}_{u,m}^{(l)}, \mathbf{z}_{u,m}^{(l-1)}). \tag{3}$$

  There are various ways to complete the update mechanism, such as GRU mechanism [14], concatenation and sum operation.
- *Readout Mechanism* is used to generate the final representations. By iteratively propagating the information in the modality graph $\mathcal{G}_m$, we obtain the nodes' representations at different layers of GNN and we use the readout function in each modality:

$$\mathbf{z}_{u,m} = f_{\text{readout}}(\{\mathbf{z}_{u,m}^{(l)}|l \in \{0, 1, \ldots, L\}), \tag{4}$$

  where L denotes the layer number of GNN. The typical way for the readout function is concatenation [7] and summation [1].

B. Multi-Modal Representation Fusion

After using GNN techniques in different modality graphs $\mathcal{G}_m$, we get representations in different modalities. Then the problem is how to fuse multi-modal representations and combine them to get the final representation. The process can be formulated as:

$$\mathbf{z}_u = f_{\text{combine}}(\{\mathbf{z}_{u,m}|m \in \mathcal{M}\}), \tag{5}$$

which can be simply set as the addition in [2] or concatenation.

C. Supervised Learning Loss

In the final prediction stage, we use the inner product to predict how user u would adopt item i:

$$\hat{y}_{ui} = \mathbf{z}_u^\top \cdot \mathbf{z}_i. \tag{6}$$

In order to find optimal parameters, we use historical interaction data to be the supervised signal. Simultaneously, due to the sparsity of data, leveraging unobserved pairs for extracting negative signals is a mainstream solution. Besides, Bayesian Personalized Ranking (BPR) loss [3] is intensively used in the recommendation.

III. Methodology

We present a multi-task learning framework for multimedia recommendation, which fuses various modal information into user and item representations and distills the multi-modal patterns. Fig. 2 illustrates the model architecture of SLMRec. Here we consider four modalities ($i.e.$, visual, acoustic, textual, and ID embeddings) serving as the basic input of the model and design three SSL tasks, which can be classified as modal-agnostic tasks and the modal-specific task. Especially, we leverage a multi-task strategy to improve the recommendation, where the graph-based recommender serves as the main task (see Section II) and SSL serves as the auxiliary task.

A. Self-Supervised Learning

Different from supervised learning, SSL is used to distill potential information in unlabelled data. Inspired by the Masked Modeling in BERT [13] and Two-Tower [17], we treat each modality as an independent feature and design two modal-agnostic tasks ($i.e.$, FD, FM). On the other hand, we consider modality specificity and relations between modalities to design a modal-specific task ($i.e.$, FAC).

1) Modal-Agnostic Task: We process features through a generic scheme ($e.g.$, masking) to generate multiple views for a node, which can be formulated as:

$$Z'_m = H(t'(E_m), \mathcal{G}_m), Z''_m = H(t''(E_m), \mathcal{G}_m), t,' t'' \sim \mathcal{T}, \tag{7}$$

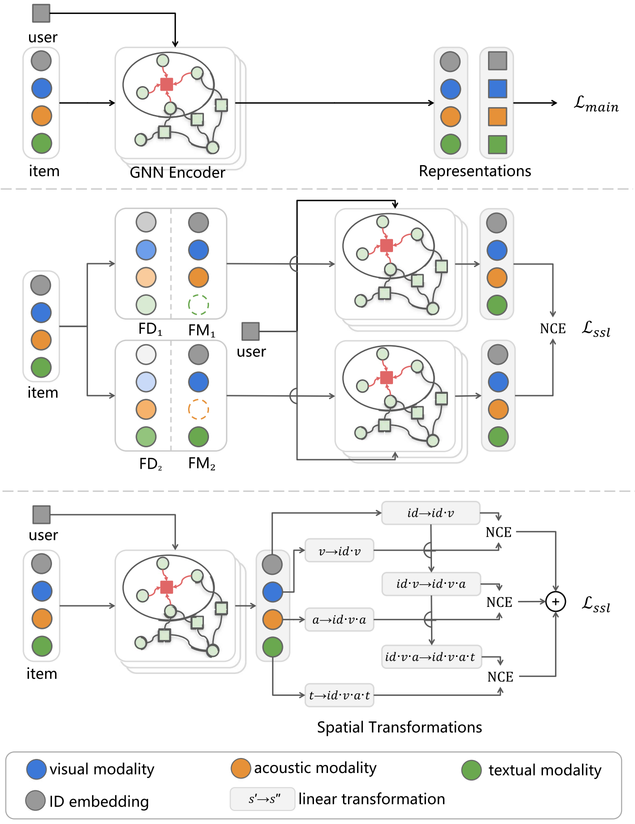


Fig. 2. The diagram depicts the main components under the multi-task strategy. The first layer represents the flow of the main task, and the next layers represent our self-supervised tasks, depicting modal-agnostic tasks and the modal-specific task, respectively.

where the mono-modal features augmented by two stochastic transformations ($i.e.$, t' and t''), from the same family of augmentations $\mathcal{T}$, are fed into the graph encoder to produce the final representations Z'_m and Z''_m. After that, for every item in each modality, we construct a positive pair defined as $(z'_{i,m}, z''_{i,m})$ and negative pairs defined as $(z'_{i,m}, z''_{j,m})$ for $i \neq j$. Especially, the modal-agnostic transformations t, which create two augmented examples by masking part of the information, can be formulated as:

$$t(E) = M \odot E, \tag{8}$$

where M and E denote the masking matrix and item features, respectively. The inner production $\odot$ of them produce the augmented example. In terms of the granularities of the masking part, we apply two masking patterns as our transformations.

- *Feature Dropout:* With a certain probability, we dropout every factor of the modal and build multiple broken views of features for contrastive learning to improve model robustness.
- *Feature Masking:* We stochastically mask one modality of items and generate two subsets of modalities. In this work, we encourage maximizing the mutual information between modalities.

Hence, we define the SSL loss on modality-agnostic tasks within the augmented embeddings:

$$\mathcal{L}_{ssl} = -\log \frac{\exp(\mathrm{sim}(z'_i, z''_i)/\tau_{ssl})}{\sum_j^N \exp(\mathrm{sim}(z'_i, z''_j)/\tau_{ssl})}, \tag{9}$$

where sim$(\cdot)$ measures the similarity between the two vectors, and cosine similarity function is applied to that; τ is a tunable hyperparameter of temperature and N denotes the number of examples in a batch.

Note that we design factor and modality granularity masking, where we call factor granularity masking as FD and modality granularity masking as FM. We provide a more specific explanation in Fig. 2. For four modalities of a given item, FD fades the node color to present the dropout process of all modalities while FM discards one node to present the masking process of a specific modality. In terms of granularity, FM further uncovers the relationship between modalities, as different modalities can be compared to maximize the mutual information that alleviates the inconsistency in modal fusion. However, conventional masking treats each modal independently and does not respect modal specificity. Next, we consider the modal-specific tasks to enforce modalities' consistency.

2) Modal-Specific Task: We generally process modalities in separate channels and fuse them through naive combinations ($e.g.$, concatenation, and addition). As we mentioned in the previous Section III-A1, intra-modal masking facilitates modal fusion but does not respect modal specificity. Hence, we hope to (1) distinguish each modality's specificity, and (2) conduct comparisons for each modality with all other modalities as much as possible for improving multi-modal recommendation. To address the above issues, we consider the treatment of different granularity spaces in multi-modal versatile network [18] to construct modal alignments based on spatial transformations.

Before alignment is performed, we encode original modality features by graph-based encoder the same as the main task. Given the mono-modal representations of items, we apply the projection head as the spatial transformation which can be defined as:

$$z_{m\to s'} = g_{m\to s'}(z_m), z_{s'\to s''} = g_{s'\to s''}(z_{m\to s'}), \tag{10}$$

where m, s', and s'' are different spaces, especially, m denotes the space S_m of each modality here. $g_{m\to s'}$ and $g_{s'\to s''}$ are simple projections to obtain the embedding $z_{m\to s'}$ and $z_{s'\to s''}$, respectively.

Considering the granularity and maneuverability of each modality, we take ID as the dominant embedding and directly align the visual, acoustic, and textual embeddings with ID embeddings in different navigable spaces respectively, which is illustrated in Fig. 2.

Following this idea, we propose a multi-layer granularity space corresponding to the different modalities: in the first level of granularity space ($S_{id\cdot v}$) the visual embeddings and the ID embeddings are compared, while in the second level of granularity space ($S_{id\cdot v\cdot a}$) the acoustic embeddings are compared with the ID embeddings and the visual embeddings, and the textual embeddings compared with the other modal embeddings in the

Algorithm 1: Multi-task Learning of SLMRec-FAC.

```
data: user-item interactions and multi-modal features
hyperparameters: $\tau, \tau_{ssl}, \alpha, \lambda$
initialization;
while not converge do
    foreach epoch do
        Perform Eq. 1 to encode modalities
        foreach batch do
            Evaluate $\mathcal{L}_{main}$ according to Eq. 13
            Perform Eq. 10 for spatial transformations
            Evaluate $\mathcal{L}_{ssl}$ according to Eq. 11
            Evaluate $\mathcal{L}$ according to Eq. 14
            Update parameters
        end
    end
end
```

last level of granularity space ($S_{id\cdot v\cdot a\cdot t}$). Note that the acoustic embeddings and visual embeddings are not compared directly. However, with the help of spatial transformation strategy that allows the visual embeddings to be projected into space $S_{id\cdot v\cdot a}$ by $g_{id\cdot v\rightarrow id\cdot v\cdot a} \circ g_{v\rightarrow id\cdot v}$ and acoustic embedding and visual embedding to be indirectly aligned based on the alignments with ID embeddings.

Hence, we define multiple InfoNCE losses [19] of direct alignments and sum them up, which can be formulated as:

$$\mathcal{L}_{ssl} = -\sum_{m\in\mathcal{M}\propto\{id\}} \log \frac{\exp(\text{sim}(z_{i,id\rightarrow s_m}, z_{i,m\rightarrow s_m})/\tau_{ssl})}{\sum_{j\in[N]}\exp(\text{sim}(z_{i,id\rightarrow s_m}, z_{j,m\rightarrow s_m})/\tau_{ssl})}, \tag{11}$$

where s_m denotes the space for contrastive learning between the modality m and ID ($e.g.$, s_a means the space $S_{id\cdot v\cdot a}$).

B. Multi-Task Strategy

To improve the recommendation result, we leverage the multi-task strategy to tackle sparse supervision signal problems and improve representation learning.

1) Contrastive Pairwise Learning: We sample one observed interaction as (u, i), and then randomly sample k items, which have not been interacted with $u(e.g.,\ j_1, j_2, \ldots, j_k)$. Thus, we establish the positive and negative pairs, which can be formulated as:

$$(u, i), (u, j_1), (u, j_2), \ldots, (u, j_k). \tag{12}$$

Here, we use LightGCN [1] as the backbone of the graph-based encoder for independently producing representations in terms of modalities. We abandon the classic BPR loss [3] as the main objective loss. Instead, we choose InfoNCE to maximize the agreement of positive pairs and minimize that of negative pairs, which can be defined as:

$$\mathcal{L}_{main} = -\log \frac{\exp\left(\text{sim}(z_u^\top z_i)/\tau\right)}{\sum_{j\in\mathcal{N}_u}\exp\left(\text{sim}(z_u^\top z_j)/\tau\right)}, \tag{13}$$

TABLE I
THE STATISTICS OF THE EVALUATION DATASET. V, A, AND T DENOTE THE FEATURE DIMENSIONS OF VISUAL, ACOUSTIC, AND TEXTUAL MODALITIES, RESPECTIVELY

Dataset	Inter.#	User.#	Item.#	V	A	T
Tiktok	564,365	36,638	71,494	256	128	128
Kwai	235,918	7,010	80,986	2,048	-	-
Movielens	1,184,023	55,485	5,986	2,048	128	100

2) Optimization: We simultaneously optimize the model by joint the main task and SSL task as followed:

$$\mathcal{L} = \mathcal{L}_{main} + \alpha\mathcal{L}_{ssl} + \lambda\|\boldsymbol{\Theta}\|_{\mathbf{2}}^{\mathbf{2}}, \tag{14}$$

where α and λ are hyperparameters to adjust the balance between multiple tasks and the strengths of L_2 regularization. We perform the multi-task optimization of SLMRec-FAC by Algorithm 1. SLMRec-FD and SLMRec-FM follow a similar workflow.

IV. EXPERIMENTS

To justify SLMRec's superiority and reveal the reasons for efficiency and effectiveness, we conduct systematic experiments to answer three research questions:

RQ1: Does SLMRec outperform the state-of-the-art models $w.r.t.$ Recall@10 for multimedia recommendation?

RQ2: Do various auxiliary tasks in SLMRec present different performances for multimedia recommendation?

RQ3: How do distinct settings influence the performance of the proposed SLMRec?

A. Experimental Settings

1) Datasets: To evaluate our devised model's efficiency, we conduct experiments on three widely used datasets, including Movielens,[2] Tiktok,[3] Kwai.[4] These datasets contain not only user-item interaction records but also collect abundant multi-modal features. Generally, all the multi-modal features have been extracted as feature vectors by the pre-training deep neural networks [2]. The statistics of all three datasets are detailed in Table I. All datasets are split into training, validation, and testing dataset with a ratio of 8:1:1 following MMGCN's dataset splitting strategy [2]. The validation datasets are used to find the optimal hyperparameters to evaluate the performance in experiments.

2) Compared Methods: We compare SLMRec with the following recommendation models. In addition to the primary matrix factorization (MF) and multi-modal recommendation ($i.e.$, VBPR and MMGCN), to highlight the performance of SLMRec for multi-modal content recommendation, we introduce the latest LightGCN for its remarkable performance.

- *MF-BPR [3]:* This method models the user and item's representation as latent vectors according to the user-item historical interactions. It then predicts whether the item fits the user's preference by measuring the similarities upon

[2] [Online]. Available: https://movielens.org/

[3] [Online]. Available: https://www.tiktok.com/

[4] [Online]. Available: https://www.kwai.com/

user and item's representations. In our experiments, we integrate the concatenation of multi-modal features and the ID embeddings as its representation.

- *LightGCN [1]:* This model is a state-of-the-art graph-based CF method. It simplifies the design of Graph Convolution Network (GCN) to improve representation learning's training efficiency and performance. In our experiment, we apply a variant of LightGCN as the backbone in SLMRec, which follows a similar structure of MMGCN to fusion the multi-modal features and ID embeddings to produce the final representation.
- *VBPR [9]:* Different from MF-BPR, this method integrates the visual features into the representation of the item in the original paper. To keep a fair comparison, our experiment concatenates the multi-modal features into one feature vector to inner-product with user's ID embeddings while devising an ID embedding and multi-modal preference vector as the user's representation to predict the interaction between users and items.
- *MMGCN [2]:* MMGCN is a well-devised model for multi-modal content recommendation. It constructs a parallel graph neural network based on user-item interactions for multiple modalities. In this model, the user's preference of the specific modality will be modeled by numerous graph convolution operations following the high-order connectivities and message-passing mechanism. MMGCN uses the same multi-modal features as SLMRec and also exploits the similar graph convolutional layers (GCN) with ours (LightGCN). Hence, the key difference between our SLMRec and MMGCN is the usage of SSL.

Furthermore, we emphasize the effectiveness of the three different SSL variants of the proposed SLMRec, shorted as SLMRec-FM, SLMRec-FD, SLMRec-FAC, which are equipped with FM, FD, and FAC, respectively.

3) Evaluation Metrics: In the evaluation phase, for each user, we view all items that the user has not interacted with as negative, while the interacted items as positive. For ranking strategy, we employ the full-rank strategy, rather than sampling a positive item and a set of negative items interacted by one or a few users. All these items are ranked based on the results of recommendation models' predictions. Moreover, we adopt Recall@K [7], Normalized Discounted Cumulative Gain (NDCG@K) [1], and Precision@K [2] as the metrics and set $K = 10$ by default, where all the evaluation metrics are widely adopted in the recommendation systems.

4) Parameter Settings: In our experiments, we program all baselines and SLMRec based on Pytorch 1.6[5] and torch-geometric package.[6] For fair comparisons, all models are initialized with Xavier and optimized by Adam [2] with the same collaborative embedding dimension of 64 and mini-batch size of 2048. In terms of the hyperparameters, we use the grid search: the learning rate and regularization weight are searched in {0.0001, 0.001, 0.01, 0.1}, where the weight-decay parameter is used as the regularization weight. For SLMRec, we search τ of the main task in {0.1, 0.2, 0.5, 1.0} by single recommendation training and fix it with 0.2. Then we tune τ_{ssl} and α within the ranges of {0.1, 0.2, 0.5, 1.0} and {0.01, 0.05, 0.1, 0.5, 1.0}, respectively. Moreover, we adopt the same stopping strategy and the same number of GNN layers as LightGCN. We do the same options for the baselines and follow their articles' designs to achieve the best performance.

TABLE II
PERFORMANCE COMPARISON BETWEEN SLMREC AND THE STATE-OF-THE-ART RECOMMENDATION ALGORITHMS ON THE THREE DATASETS. THE %IMPROV. DENOTES THE RELATIVE PERFORMANCE IMPROVEMENT OVER STRONGEST BASELINE

Metric	Model	Tiktok	Kwai	Movielens
Recall@10	MF-BPR	0.073985	0.039913	0.130689
	LightGCN	0.074568	0.046371	0.135659
	VBPR	0.022105	0.019401	0.156711
	MMGCN	0.026078	0.024481	0.135288
	SLMRec-FAC	**0.132955**	**0.057304**	**0.270616**
	%Improv.	78.30%	23.58%	72.68%
NDCG@10	MF-BPR	0.042571	0.034357	0.086055
	LightGCN	0.043207	0.039368	0.088493
	VBPR	0.012653	0.016082	0.102065
	MMGCN	0.018451	0.010931	0.087006
	SLMRec-FAC	**0.078081**	**0.045535**	**0.186821**
	%Improv.	80.71%	15.67%	83.04%
Precision@10	MF-BPR	0.011892	0.010428	0.010571
	LightGCN	0.012201	0.013309	0.033536
	VBPR	0.002787	0.004265	0.037528
	MMGCN	0.004514	0.006191	0.031240
	SLMRec-FAC	**0.022694**	**0.019301**	**0.065286**
	%Improv.	86.000%	45.02%	73.97%

B. Performance Comparison

1) Comparison With the State-of-The-Arts: As shown in Table II, we summarize all experiments' results and report the improvements calculated between SLMRec and the best performance baselines highlighted with an underline. And we have the following observations:

- In all evaluation metrics of Recall@10, NDCG@10, and Precision@10, SLMRec's performance outperforms all the state-of-the-art baselines across all the datasets. Without any doubt, the results fully present the rationality and superiority of our method and answer **RQ1** powerfully. In particular, SLMRec makes the improvements over the strongest baseline $w.r.t.$ Recall@10 by 78.30%, 23.58%, and 72.68% on the three datasets. We attribute its superiority to the following aspects: 1) the specially designed architecture of SLMRec with contrastive loss helps to learn the collaborative information, and 2) the auxiliary SSL component helps to enhance model robustness and capture mutual information between different modalities.
- By analyzing all the datasets' results, we find that SLMRec achieves more significant promotion on the Tiktok dataset than others. It is reasonable since we only adopt visual features on the Kwai dataset, while the Movielens dataset is less sparse with much fewer items. On the other hand, it demonstrates our devised SSL model has better capability to distill high-level semantic information through multi-modal in the sparse space.

[5] [Online]. Available: https://pytorch.org/

[6] [Online]. Available: https://pytorch-geometric.readthedocs.io/

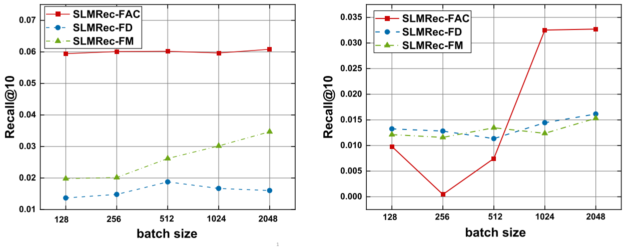
TABLE III
PERFORMANCE COMPARISON OF SLMREC AND BASELINES (LIGHTGCN IN TIKTOK AND KWAI, VBPR IN MOVIELENS)

Metric	Model	Tiktok	Kwai	Movielens
Recall@10	Baseline	0.074568	0.046371	0.156711
	SLMRec-FM	0.131209	0.057006	0.269497
	SLMRec-FD	0.130143	0.057060	0.268943
	SLMRec-FM+FD	0.130804	0.057236	0.268208
	SLMRec-FAC	**0.132955**	**0.057304**	**0.270616**
NDCG@10	Baseline	0.043207	0.039368	0.102065
	SLMRec-FM	0.076837	0.045002	0.185904
	SLMRec-FD	0.076662	0.045517	0.185865
	SLMRec-FM+FD	0.076732	0.044929	0.185251
	SLMRec-FAC	**0.078081**	**0.045535**	**0.186821**
Precision@10	Baseline	0.012201	0.013309	0.037528
	SLMRec-FM	0.022413	0.018802	0.064937
	SLMRec-FD	0.022187	0.019101	0.064975
	SLMRec-FM+FD	0.022268	0.019101	0.064169
	SLMRec-FAC	**0.022694**	**0.019301**	**0.065286**

- LightGCN outperforms other baselines on Tiktok and Kwai Datasets. Since GNN-based recommenders help to model user and item's representations upon high-order connectivities with message passing mechanism. Especially in sparse data, the advantages of GNN-based recommenders are more obvious.
- We notice that VBPR achieves better performance on the Movielens dataset than all baselines, while underperforms other methods significantly on the Tiktok and Kwai datasets. The reason may be VBPR adopts a simple design with the preference vector, which is more good at less multi-modal features and intensive interactions. Therefore, when the number of items is minor, VBPR shows much better performance but becomes unstable when the items' amount increases.
- Usually, we percept that the performance of the graph-based model with introducing multi-modal information should be better. However, we observe that MMGCN shows worse performance than MF-BPR in Tiktok and Kwai such sparsy data, and VBPR in Movielens. It may be that its unreasonable design, mentioned in LightGCN, injects much noise and affects the performance. Besides, the Tiktok dataset and Kwai dataset are more sparse than Movielens. The sparsity of datasets may make the model training insufficient. It demonstrates the reason that ranking in a set of samples, in the original MMGCN paper, archives better performance than the all-ranking strategy adopted in our experiments. Moreover, as we mentioned before, the key difference between our SLMRec and MMGCN is the usage of SSL, so it further verifies the superiority of the SSL.

We conduct relative experiments on modal-agnostic and model-specific auxiliary tasks and present the result comparison to answer **RQ2**. As shown in Table III, we find that:

- All SLMRec models outperform the strongest performance significantly. It further justifies bringing the SSL's superiority into recommendation representation learning facilitates capturing latent patterns among modalities.



(a) Tiktok

(b) Kwai

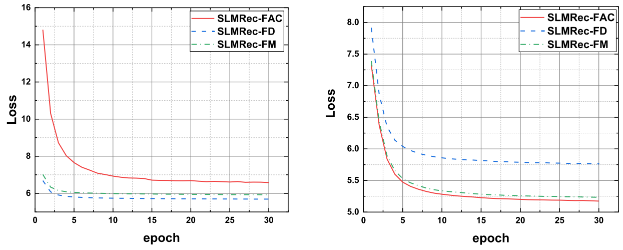
Fig. 3. Effect of SSL tasks $w.r.t.$ Recall@10.

- SLMRec-FAC outperforms SLMRec-FM and SLMRec-FD relatively on all three datasets. It demonstrates complying with the modal-specific and considering fine and coarse features benefit representation learning for multimedia recommendation. In other words, adopting modalities alignments on navigable spaces in SLMRec-FAC distills fine and coarse multi-modal patterns.
- SLMRec-FM shows better result than SLMRec-FD in Tiktok but achieve similar performance on the other two datasets. The two methods are similar in structure technically while with the different granularity of the masking part. Therefore, the advantage of SLMRec-FM is diminished when faced with the datasets of few modalities (i.e., Kwai) and the dense interaction (i.e., Movielens).
- The combination of FM and FD is an extension of the masking part on representation vectors, wherein FM is for modal-grain and FD is for factor-grain. SLMRec-FM+FD is not significantly better than FM or FD, which can be attributed to SLMRec-FM+FD being another variant of modal-agnostic task with broader masking part, thereby limiting its breakthrough.
- The performances of all models are similar on the Kwai dataset. It is reasonable since we only adopt visual modal on the Kwai dataset.

C. Ablation Study

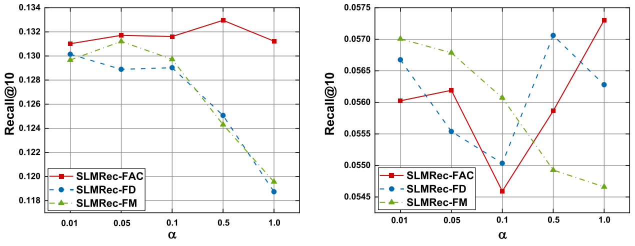
Affected by the primary task, the gap in the experimental results is not significant in the previous section. In this section, we present ablation studies for mining deep insights of SLMRec to further address **RQ2**. Hence, we discard the main task while leaving SSL tasks to optimize the model and evaluate by the same operations as multi-task. And we find that the SSL task in SLMRec-FAC can significantly assist the main task to optimize the model with a large value of α ($e.g.$, 1.0), which will be explored in Section IV-D2. Fig. 3 shows the performance of the FAC task discarding the main task $w.r.t.$ Recall@10 in Tiktok and Kwai on the range of batch size in {128, 256, 512, 1024, 2048} where larger batch size denotes larger negative samples for contrastive learning. Here, we get the following four observations:

- SLMRec-FAC still performs well without the main task ($i.e.$, 0.0608 in Tiktok and 0.032716 in Kwai $w.r.t.$ Recall@10) and even surpasses some of the baselines ($e.g.$,



(a) Tiktok (b) Kwai

Fig. 4. Training curves of SLMRec.



(a) Tiktok (b) Kwai

Fig. 5. Impact of regularization strength α for SLMRec.

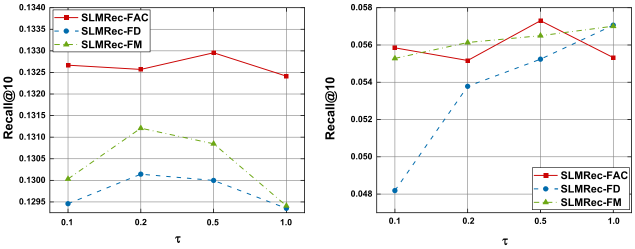
MF, VBPR) on both evaluated datasets checking on Table II, which further demonstrates its superiority for recommendation leveraging multi-modal information.

- SLMRec-FAC far exceeds SLMRec-FD and SLMRec-FM in performance. we speculate that modality alignment based on multi-level granularities can better extract key factors in modalities and construct more effective representations for recommendation than direct comparison between multiple views of original data.
- Enlarge value of batch size increases the results, which intuitively can be attributed to more negative samples compared with the positive sample to distinguish the differences and improve representation learning. However, in the variation of batch size, SLMRec-FAC shows more stability in Tiktok than in Kwai. One possible reason is that SLMRec-FAC depends more on modalities than multiple negative pairs.
- SLMRec-FAC is unstable with small batch sizes in Kwai. Since SLMRec-FAC is a modal-specific method and Kwai only contains visual features, the smaller batch size may limit the contrastive learning of SLMRec-FAC and the fewer modalities would make the performance of SLMRec-FAC unstable. As a result, when the batch size is 256, SLMRec-FAC is stocking in a suboptimal status at the first few epochs and never improves.

D. Study of SLMRec

In this section, we will perform three experiments to answer **RQ3**. In particular, we assume all auxiliary tasks as settings, to study the effect on the convergence rate of SLMRec.

1) Training Efficiency: Fig. 4 shows the loss convergence of SLMRec family in Tiktok and Kwai, while the loss curve smoothes out, the model will gradually converge. We find that SLMRec-FAC is obviously slower to converge than the other two methods in Tiktok the contrary to the similar convergence speeds of all methods on the Kwai dataset. SLMRec-FAC constructs multi-layer granularities spaces and builds modalities alignments to enforce mutual information of modalities, especially in a modal-rich dataset ($e.g.$, Tiktok) but takes more time to converge. Notably, on the Kwai dataset with a single modality, SLMRec-FAC only needs to align ID embeddings with one modality, so it has similar convergence speed and performance (see Table III) as SLMRec-FD and SLMRec-FM.



(a) Tiktok (b) Kwai

Fig. 6. Impact of temperature hyperparameter τ for SLMRec.

2) Effect of α: Fig. 5 shows performance comparisons evaluated by SLMRec-FAC, SLMRec-FD, and SLMRec-FM $w.r.t.$ Recall@10 under different α in Tiktok and Kwai. It shows that FD and FM tasks are more dependent on the main task than FAC, while high-weighted SLMRec-FD and SLMRec-FM lead to worse results, especially in Tiktok. Hence, we intuitively speculate that FAC can be applied as a task for multi-modal recommendation learning because an enlarged value of α brings better performance in the SLMRec-FAC task, and Section IV-C demonstrates the performance of SLMRec-FAC discarding the main task.

3) Effect of τ: τ controls the scale of InfoNCE and the smaller τ amplifies error correction capability while the bigger τ enhances fault tolerance. Fig. 6 indicates that SLMRec-FAC has stable performance with $\tau = 0.5$ in Tiktok and Kwai while SLMRec-FD and SLMRec-FM need fine-grained tuning of τ to obtain the best performance that the large τ leads performance drop in Tiktok but brings good performance in Kwai.

V. Related Work

In this section, we introduce some works that are related to our method, including multi-modal personalized recommendation and self-supervised learning.

A. Multi-Modal Personalized Recommendation

In Recommendation Systems, collaborative filtering (CF) [4], [5], [7], [20], [21] models are commonly used for personalized tasks. CF-based approaches leverage historic feedback to learn a low-dimensional vector to capture user preference. However, due to the highly-skewed data distribution in historic interactions [17], [22], CF-based approaches work poorly when meeting the item with few interactions ($i.e.$ cold-start problem).

To alleviate the cold-start problem, researchers have tried various methods, one of which is to develop hybrid approaches that incorporate item's side information [23], [24]. For instance, He $et\ al.$ [9] proposed a scalable factorization model, short as VBPR, to integrate visual information for predicting users' opinions. Guan $et\ al.$ [25] utilized the underlying community structure to regularize user latent preferences. Gao $et\ al.$ [26] applied the attention mechanism to learn user preference through the recipe's visual feature. Besides, Barkan $et\ al.$ [27] uses Mean Squared Error to narrow the distance between CF-based representation and CB-based representation. A series of work [2], [28]–[31] argued that the prior work ignores the differences and user preferences upon modalities then fail to model the user preference on different modalities. Instead, they construct the graph in each modality and corporate the multi-modal features into the item's representation. Alayrac $et\ al.$ [18] explores how best to combine the visual, acoustic, and textual modalities, and considers three options for the modality embedding graphs: shared space, disjoint space, fine and coarse space (FAC). Enlightened by MMGCN [2] and Alayrac $et\ al.$ [18], we split the user-item graph by modalities and design the self-supervised auxiliary task as FAC in our model.

B. Self-Supervised Learning

Over the years, self-supervised learning achieved remarkable successes in many domains. In CV, a large collection of work [11], [12] employ contrastive learning [32] to learn a better representation by defining pretext task as an instance discrimination task [33]. More specifically, contrastive learning compares different views of the data and distinguishes whether samples are transformed from the same data. The goal of contrastive learning is to maximize the mutual information between different views. To achieve it, Oord $et\ al.$ [19] proposed a probabilistic contrastive loss, called InfoNCE. Oord also pointed that minimizing the InfoNCE loss maximizes a lower bound on mutual information. Yuan $et\ al.$ [34] generate pseudo-labels of consecutive video frames and utilize the proposed multi-cycle consistency loss to learn a feature extraction network. In NLP, BERT [13] uses masked-LM and next sentence prediction as pre-training tasks and gets empirical improvement.

There are many researches applying SSL to graph data. For example, Hu $et\ al.$ [35] develop a new strategy and self-supervised methods for pre-training GNNs. GCC [36] leverages contrastive learning to learn the transferable structural representations. Hu $et\ al.$ [37] designs pre-training tasks for GNNs to extract the generic graph structure information. You $et\ al.$ [38] analyze how to incorporate self-supervised learning in GCNs and demonstrate the superiority of multi-task learning over pre-training and self-training.

Meanwhile, the application of SSL in the recommendation system is still less explored [39]. Yao $et\ al.$ [17] adopts the multi-task scheme with SSL and takes two-tower DNN as encoder, while in sequential recommendation, inspired by the framework of SimCLR [11] and Yao $et\ al.$ [17], CP4Rec [40] borrows a similar idea of contrastive pre-training framework while S^3-Rec [41] utilizes the intrinsic data correlation to derive self-supervision signals. Besides, Xin $et\ al.$ [42] proposes self-supervised reinforcement learning that augments standard recommendation models with a self-supervised learning output layer and a reinforcement learning output layer.

VI. Conclusion and Future Work

In this research, we presented current multimedia recommender models that mostly followed a supervised learning paradigm, which suffered from sparse supervisory data and untouched patterns and structures inherent, and explored the potential of SSL to solve the limitations. We devised a parallel graph-based recommender model, served as the primary supervised learning task, and a multi-modal SSL component as the auxiliary task. From the patterns of multiple modalities, we devised three types of data augmentation at different granularity to construct the auxiliary task. We conducted extensive experiments on three benchmark datasets to justify the superiority of our proposed model. Empirical results showed that SLMRec exhibited substantial improvements over the state-of-the-art baselines.

In this work, we attempt to incorporate self-supervised learning for multimedia recommendation. In future work, we will exploit more powerful data augmentation methods to capture multi-modal patterns. Meanwhile, data augmentations based on graph structure would also be introduced, such as Node Dropout (ND), Edge Dropout (ED). Moreover, we observe a significant effect from the bias of data, and we will try to utilize the causality [43], [44] to improve the recommendation performance for multimedia.

References

[1] X. K. He, X. Deng, Y. W. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM Conf. Res. Develop. Inf. Retrieval*, J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. Virtual Event, China, ACM, Jul. 2020, pp. 639–648.

[2] Y. Wei et al., "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proc. 27th ACM Int. Conf. Multimedia*, L. Amsaleg, B. Huet, M. A. Larson, G. Gravier, H. Hung, C. Ngo, and W. T. Ooi, Eds. Nice, France, ACM, Oct. 2019, pp. 1437–1445.

[3] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell.*, J. A. Bilmes and A. Y. Ng, Eds. Montreal, QC, Canada, AUAI Press, Jun. 2009, pp. 452–461.

[4] X. He et al., "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, Eds. Perth, Australia, ACM, Apr. 2017, pp. 173–182.

[5] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proc. 14th ACM Int. Conf. Knowl. Discov. Data Mining*, Y. Li, B. Liu, and S. Sarawagi, Eds. Las Vegas, Nevada, USA, ACM, Aug. 2008, pp. 426–434.

[6] F. Xue et al., "Deep item-based collaborative filtering for top-n recommendation," *TOIS*, vol. 37, no. 3, pp. 1–25, 2019.

[7] X. Wang, X. He, M. Wang, F. Feng, and T. Chua, "Neural graph collaborative filtering," in *Proc. 42nd Int. ACM Conf. Res. Develop. Inf. Retrieval*, B. Piwowarski et al., Eds. Paris, France, ACM, Jul. 2019, pp. 165–174.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, IEEE Computer Society, 2016, pp. 770–778.

[9] R. He and J. J. McAuley, “VBPR: Visual bayesian personalized ranking from implicit feedback,” in *Proc. 30th Conf. Artif. Intell.*, D. Schuurmans and M. P. Wellman, Eds. Phoenix, Arizona, USA, AAAI Press, Feb. 2016, pp. 144–150.

[10] J. Chen et al., “Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention,” in *Proc. 40th Int. ACM Conf. Res. Develop. Inf. Retrieval*, N. Kando et al., Eds., Shinjuku, Tokyo, Japan, Aug. 2017, pp. 335–344.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. 37th Int. Conf. Mach. Learn., 2020*, 2020, vol. 119, pp. 1597–1607.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 9729–9738.

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, MN, USA, Association for Computational Linguistics, Jun. 2019, vol. 1, pp. 4171–4186.

[14] Y. Li, D. Tarlow, M. Brockschmidt, and R. S. Zemel, “Gated graph sequence neural networks,” in *Proc. 4th Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds. San Juan, Puerto Rico, May 2016.

[15] P. Velickovic et al., “Graph attention networks,” in *Proc. 6th Int. Conf. Learn. Representations*, Vancouver, BC, Canada, 2018.

[16] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, “Graph neural networks in recommender systems: A survey,” *ACM Comput. Surv.*, to be published, doi: 10.1145/3535101.

[17] T. Yao et al., “Self-supervised learning for deep models in recommendations,” in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 4321–4330.

[18] J. Alayrac et al., “Self-supervised multimodal versatile networks,” in *Proc. Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst.*, virtual, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Dec. 2020, pp. 25–37.

[19] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018, *arXiv:1807.03748*.

[20] S. Rendle, “Factorization machines,” in *Proc. 10th IEEE Int. Conf. Data Mining*, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, Eds. Sydney, Australia, IEEE Computer Society, Dec. 2010, pp. 995–1000.

[21] X. Wang et al., “Item silk road: Recommending items from information domains to social users,” in *Proc. 40th Int. ACM Conf. Res. Develop. Inf. Retrieval*, N. Kando et al., Eds. Shinjuku, Tokyo, Japan, ACM, Aug. 2017, pp. 185–194.

[22] Z. Tao, X. Wang, X. He, X. Huang, and T. Chua, “Hoafm: A high-order attentive factorization machine for CTR prediction,” *Inf. Process. Manag.*, vol. 57, no. 6, 2020, Art. no. 102076.

[23] N. Xu et al., “Multi-level policy and reward-based deep reinforcement learning framework for image captioning,” *IEEE Trans. Multim.*, vol. 22, no. 5, pp. 1372–1383, May 2020.

[24] A. Liu et al., “Adaptively clustering-driven learning for visual relationship detection,” *IEEE Trans. Multim.*, vol. 23, pp. 4515–4525, 2021.

[25] J. Guan, X. Huang, and B. Chen, “Community-aware social recommendation: A unified SCSVD framework,” *IEEE Trans. Knowl. Data Eng.*, to be published, doi: 10.1109/TKDE.2021.3117686.

[26] Y. Wei et al., “Hierarchical user intent graph network for multimedia recommendation,” *IEEE Trans. Multimedia*, vol. 24, pp. 2701–2712, 2021.

[27] O. Barkan, N. Koenigstein, E. Yogev, and O. Katz, “CB2CF: A neural multiview content-to-collaborative filtering model for completely cold item recommendations,” in *Proc. 13th ACM Conf. Recommender Syst.*, T. Bogers, A. Said, P. Brusilovsky, and D. Tikk, Eds. Copenhagen, Denmark, ACM, Sep. 2019, pp. 228–236.

[28] Z. Tao et al., “MGAT: Multimodal graph attention network for recommendation,” *Inf. Process. Manag.*, vol. 57, no. 5, 2020, Art. no. 102277.

[29] F. Xue et al., “Knowledge-based topic model for multi-modal social event analysis,” *IEEE Trans. Multimedia*, vol. 22, no. 8, pp. 2098–2110, Aug. 2020.

[30] X. Chen, D. Liu, Z. Xiong, and Z. Zha, “Learning and fusing multiple user interest representations for micro-video and movie recommendations,” *IEEE Trans. Multimedia*, vol. 23, pp. 484–496, 2021.

[31] J. Zahálka, S. Rudinac, and M. Worring, “Interactive multimodal learning for venue recommendation,” *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2235–2244, Dec. 2015.

[32] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA. IEEE Computer Society, 2006, pp. 1735–1742.

[33] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, IEEE Computer Society, 2018, pp. 3733–3742.

[34] D. Yuan, X. Chang, P.-Y. Huang, Q. Liu, and Z. He, “Self-supervised deep correlation tracking,” *IEEE Trans. Image Process.*, vol. 30, pp. 976–985, 2021.

[35] W. Hu et al., “Strategies for pre-training graph neural networks,” in *Proc. 8th Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=HJlWWJSFDH

[36] J. Qiu et al., “GCC: Graph contrastive coding for graph neural network pre-training,” in *Proc. 26th ACM Conf. Knowl. Discov. Data Mining*, R. Gupta, Y. Liu, J. Tang, and B. A. Prakash, Eds. Virtual Event, CA, USA, ACM, Aug. 2020, pp. 1150–1160.

[37] Z. Hu, C. Fan, T. Chen, K.-W. Chang, and Y. Sun, “Pre-training graph neural networks for generic structural feature extraction,” 2019, *arXiv:1905.13728*.

[38] Y. You, T. Chen, Z. Wang, and Y. Shen, “When does self-supervision help graph convolutional networks?,” in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 10871–10880.

[39] J. Wu et al., “Self-supervised graph learning for recommendation,” in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 726–735.

[40] X. Xie et al., “Contrastive learning for sequential recommendation,” 2020, *arXiv:2010.14395*.

[41] K. Zhou et al., “S3-REC: Self-supervised learning for sequential recommendation with mutual information maximization,” in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, M. d’Aquin, S. Dietze, C. Hauff, E. Curry, and P. Cudré-Mauroux, Eds. Virtual Event, Ireland, ACM, Oct. 2020, pp. 1893–1902.

[42] X. Xin, A. Karatzoglou, I. Arapakis, and J. M. Jose, “Self-supervised reinforcement learning for recommender systems,” in *Proc. 43 rd Int. ACM Conf. Res. Develop. Inf. Retrieval*, J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, and Y. Liu, Eds. Virtual Event, China, ACM, Jul. 2020, pp. 931–940.

[43] Y. Wu, X. Wang, A. Zhang, X. He, and T. S. Chua, “Discovering invariant rationales for graph neural networks,” in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 2928–2937.

[44] Y. Li, X. Wang, J. Xiao, W. Ji, and T. S. Chua, “Invariant grounding for video question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2928–2937.