



Semantic-Guided Feature Distillation for Multimodal Recommendation

Fan Liu
National University of Singapore
liufancs@gmail.com

Huilin Chen
Hefei University of Technology
ClownClumsy@outlook.com

Zhiyong Cheng
Qilu University of Technology
(Shandong Academy of Sciences)
jason.zy.cheng@gmail.com

Liqiang Nie
Harbin Institute of Technology,
Shenzhen
nieliqiang@gmail.com

Mohan Kankanhalli
National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

Multimodal recommendation exploits the rich multimodal information associated with users or items to enhance the representation learning for better performance. In these methods, end-to-end feature extractors (e.g., shallow/deep neural networks) are often adopted to tailor the generic multimodal features that are extracted from raw data by pre-trained models for recommendation. However, compact extractors, such as shallow neural networks, may find it challenging to extract effective information from complex and high-dimensional generic modality features. Conversely, DNN-based extractors may encounter the data sparsity problem in recommendation. To address this problem, we propose a novel model-agnostic approach called Semantic-guided Feature Distillation (SGFD), which employs a teacher-student framework to extract feature for multimodal recommendation. The teacher model first extracts rich modality features from the generic modality feature by considering both the semantic information of items and the complementary information of multiple modalities. SGFD then utilizes response-based and feature-based distillation loss to effectively transfer the knowledge encoded in the teacher model to the student model. To evaluate the effectiveness of our SGFD, we integrate SGFD into three backbone multimodal recommendation models. Extensive experiments on three public real-world datasets demonstrate that SGFD-enhanced models can achieve substantial improvement over their counterparts.

CCS CONCEPTS

• **Information systems** → **Personalization; Recommender systems; Collaborative filtering.**

KEYWORDS

Multimodal Recommendation, Knowledge Distillation, Deep Learning, Collaborative Filtering

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10...\$15.00
<https://doi.org/10.1145/3581783.3611886>

ACM Reference Format:

Fan Liu, Huilin Chen, Zhiyong Cheng, Liqiang Nie, and Mohan Kankanhalli. 2023. Semantic-Guided Feature Distillation for Multimodal Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3581783.3611886>

1 INTRODUCTION

Recommendation system plays an essential role in various online platforms, such as E-commerce, social media, and streaming services. Many recommendation approaches are based on Collaborative Filtering (CF) [2, 8, 13, 14, 20, 23, 25, 38], which efficiently leverages user-item interactions. Nevertheless, the CF-based recommendation model is incapable of accurately capturing user preferences when the interaction data are sparse, resulting in poor recommendation performance. To tackle the limitation, many methods have been developed to leverage the side information, such as attributes, reviews, and images, associated with users and items, which provides rich information about user preference and item characteristics [5, 11, 24, 27, 50]. Recognizing that features from different modalities can be complementary to each other, various approaches have been proposed to exploit multimodal features for learning more comprehensive user and item representations for recommendation [16, 21, 22, 39, 42–44, 49, 51].

Extracting effective features from multimodal data serves as a foundation for a multimodal recommendation model. Fig. 1 uses the visual feature extraction process to illustrate the general pipeline for feature extraction in the multimodal recommendation methods. Typically, pre-trained models (such as VGG [30] for visual features and Bert [10] for textual features) are firstly adopted to extract generic features, which are then tailored for recommendation via a designed feature extractor (e.g., shallow/deep neural network). The multimodal features distilled from the feature extractor are subsequently used as input in the recommendation model to learn user and item representations. For example, Liu et al. [22] presented a metric learning-based method that models the diverse preferences of users by exploiting the multimodal features of the item to estimate the special attention of the user to different aspects of that item. To prune potential false-positive edges in the user-item interaction graph, Wei et al. [44] designed a method that scores the affinity between user preference and multimodal features of an item to measure the confidence of edge being true-positive interaction. It is worth mentioning that, in those methods, the feature extractor

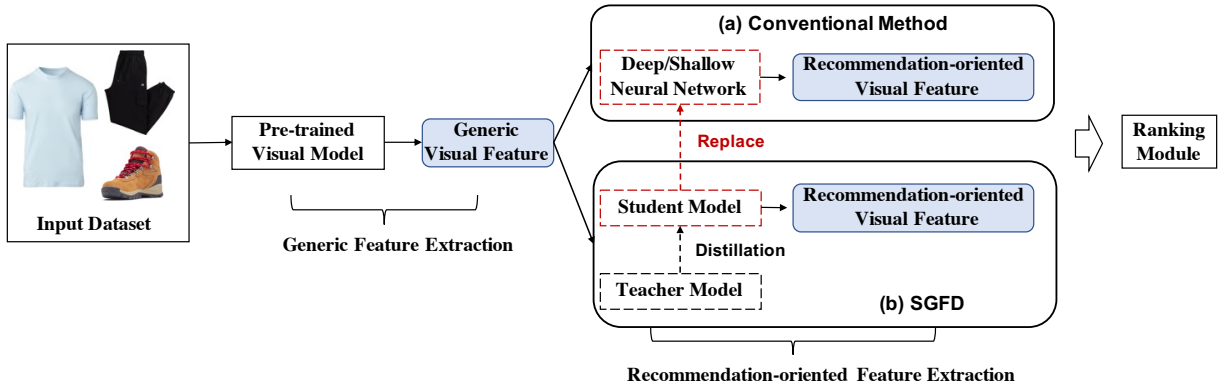


Figure 1: An illustration of the feature extraction pipeline in multimodal recommendation models. (a) is the conventional feature extraction method and (b) is our proposed SGFD.

is often jointly learned with the user and item representations in the recommendation model in an end-to-end manner.

Despite the progress, many existing methods overlooked the importance of feature extraction in multimodal recommendation. Due to the complexity and high dimensionality of multi-modal features, a compact feature extractor adopting shallow neural networks (SNNs) can hardly extract effective information from the generic features of each modality. Therefore, most recent multimodal recommendation models employ deep neural networks as the extractor [16, 22, 42, 44, 49], leveraging the powerful representation learning capability of DNNs. However, the problem is that deep models need sufficient training data for good performance. Unfortunately, recommendation models often face serious data sparsity problem. As a result, feature extractors with deep models cannot be well trained, which not only renders them unhelpful but can also negatively impact the final recommendation performance. Therefore, the primary challenge faced by existing multimodal recommendation methods is the dilemma of choosing between deep or shallow models as feature extractors, which is an essential part of the multimodal recommendation model. The second challenge is that semantic information is often ignored during the feature extraction process. In fact, existing recommendation models extract modality features only using collaborative information [22, 44, 51]. Yet, semantic information contains rich semantic and concise knowledge, which can well guide the feature extraction process. For example, category information is a more abstract and general form of semantic information, and it can help recommendation models identify relationships or similarities between items.

Inspired by the above considerations, in this paper, we introduce a novel model-agnostic approach, termed **Semantic-Guided Feature Distillation (SGFD for short)**, which can robustly extract effective recommendation-oriented features from generic multimodal features using the teacher-student framework. Specifically, we propose a teacher model composed of feature extractors with deep neural networks for different modalities. These feature extractors are designed to extract features considering both the semantic information of items and the complementary information of multiple modalities. As the teacher model is separately trained with

recommendation model, it will not suffer the data sparsity problem in recommendation. Meanwhile, we use the feature extractors with shallow neural networks as the student model to replace the original feature extractors in recommendation methods. In addition, we introduce a feature-based distillation loss and a response-based distillation loss to transfer knowledge from the teacher model to the student model. To evaluate the efficacy of our approach, we conducted comprehensive experiments on three real-world datasets. The results demonstrate that our method can significantly outperform existing multimodal recommendation models. We have released the code and relevant parameter settings to facilitate repeatability as well as further research¹.

In summary, the main contributions of this work are as follows:

- We emphasize the constraints of existing recommendation methods in effectively utilizing multimodal features. To overcome these limitations, we propose a model-agnostic approach SGFD that can extract effective recommendation-oriented features from generic modality features using the teacher-student framework.
- To leverage the semantic information during the feature extraction processes, we design feature extractors that consider both the semantic information of items and the complementary information of multiple modalities.
- We employ the response-based and feature-based distillation loss to facilitate knowledge transfer from the teacher to the student model.
- We have conducted extensive experiments on three real-world datasets, demonstrating the superiority of our method.

2 RELATED WORK

2.1 Multimodal Recommendation

Various types of side information have been utilized in recommender systems, such as attributes [5, 24], reviews [3, 4, 6, 7, 9, 26, 33] and images [11, 17, 27, 37]. Most existing models exploit different types of information individually rather than fusing them, due to the challenges involved in effectively combining different

¹<https://github.com/HuilinChenJN/SGFD>.

modalities of information. However, it is well-recognized that multimodal features can provide complementary information to each other, as demonstrated in [22, 31, 32, 36, 47, 49]. Thus, various approaches have been proposed to learn the user and item representations by fusing multimodal features. For example, JRL [49] first extracts user and item features from ratings, reviews, and images separately with deep neural networks, then concatenates those features to form the final user and item representations. In recent years, Graph Convolution Networks (GCNs) have attracted increasing attention in multimedia recommendation due to their powerful capability on representation learning from non-Euclidean structure [42, 44, 48]. MMGCN [42] learns the model-specific user preference to the content information via the direct information interchange between user and item in each modality. For reducing the computational cost on large graphs and removing noise supervision signals caused by the negative sampling strategy, BM3 [51] introduces a self-supervised method which can learn the representations of users and items by jointly optimizing three multimodal objectives.

Despite the progress, the feature extractor in these methods is jointly optimized with the CF-based recommendation approach. However, a compact extractor may not extract effective modality features, and a DNN-based extractor suffers from data sparsity in recommendation. In this paper, to utilize more useful multimodal information for user and item representation learning, we employ the knowledge distillation technique to extract more effective recommendation-oriented multimodal features from generic multimodal features.

2.2 Knowledge Distillation

Knowledge Distillation (KD) is composed of a “compact” model (i.e., student model) and a previously trained “large” model (i.e., teacher model) [1, 15, 19, 28]. The core of KD is that a previously trained large teacher model transfers knowledge to a small student model, so that the student model performs comparably to the teacher model. Moreover, the student model also has lower inference latency due to its small size and compact network architecture. Feature Distillation (FD) is a primary method in KD, which distill features in the teacher network through a transformation [1, 29, 35].

Recently, inspired by the success of KD in various fields, it has been introduced into the Recommendation System (RS) to enhance the capability of modeling users’ preferences or reduce inference latency while maintaining performance [18, 34, 45], such as Ranking Distillation (RD), Collaborative Distillation (CD). Ranking Distillation (RD) learns to rank documents/items from both the training data and the supervision of a larger teacher model [18, 34]. For example, Kang et al. [18] proposed a method that enables the student model to learn from the latent knowledge encoded in the teacher model and from the teacher’s predictions results. Collaborative Distillation (CD) first samples unobserved items from the teacher’s recommendation list according to their ranking, then trains the student to mimic the teacher’s prediction score (e.g., relevance probability) on the sampled items [45].

In this paper, we propose a teacher model which extracts multimodal features from the generic feature, guided by semantic information. Meanwhile, we employ a student model with shallow

neural networks to serve as feature extractors for various modalities in recommendation model.

3 OUR METHOD

3.1 Preliminaries

3.1.1 Task Formulation. In this paper, we focus on feature extraction in multimodal recommendation. To alleviate the data sparsity problem in recommendation, multimodal features (e.g., user review and item image²) are leveraged to learn better representations of users and items. In multimodal recommendation, the generic review and image features are extracted from multimodal data using pre-trained models. The main objective of our work is to derive effective recommendation-oriented features from these generic multimodal features to improve recommendation accuracy. In other words, our task is to effectively utilize the information contained in the reviews and images of items to better capture users’ preferences and make more accurate recommendations.

3.1.2 Generic Feature Extraction. In our implementation, we use the VGG16 model [30] to extract generic visual features from item images. This model consists of 13 convolutional layers and three fully connected layers. It is trained by using 1.2 million ImageNet (ILSVRC2010) images. We utilize the output from the second fully connected layer, a 4096-dimensional feature vector, as the generic visual feature F_m . Additionally, we employ BERT [10] to extract generic textual features. This model is capable of deriving deep bidirectional representations from the unlabeled text by joint conditioning on both left and right contexts across all layers. In our work, we input all the users’ reviews for an item to BERT to obtain the generic textual feature F_e for this item.

3.1.3 Recommendation-oriented Feature Extraction. After obtaining the generic visual and textual features F_m and F_e , the feature extractors in the multimodal recommendation model are then used to tailor these features for recommendation. Take visual feature as an example, previous studies [22, 44, 51] feed the generic visual feature F_m into a Multi-Layer Perceptrons (MLP) to extract the recommendation-oriented visual feature f_m . Specifically, it is defined as:

$$f_m = \text{MLP}(F_m). \quad (1)$$

It should be noted that the adopted $\text{MLP}(\cdot)$ are jointly optimized with the recommendation model. Thus, the training of feature extractors will also suffer from the data sparsity problem in recommendation. The recommendation-oriented textual feature f_s can be obtained similarly. In multimodal recommendation models, the textual feature F_e and visual feature F_m will be fed into two separate MLPs to extract the effective recommendation-oriented features f_e and f_m , respectively. These extracted features f_e and f_m are subsequently used to improve recommendation accuracy.

3.2 Semantic-guided Feature Distillation

We propose a model-agnostic approach, termed semantic-guided feature distillation, to extract effective recommendation-oriented

²Note that only the multimodal data of items is used in this work. Besides, each item is associated with one image.

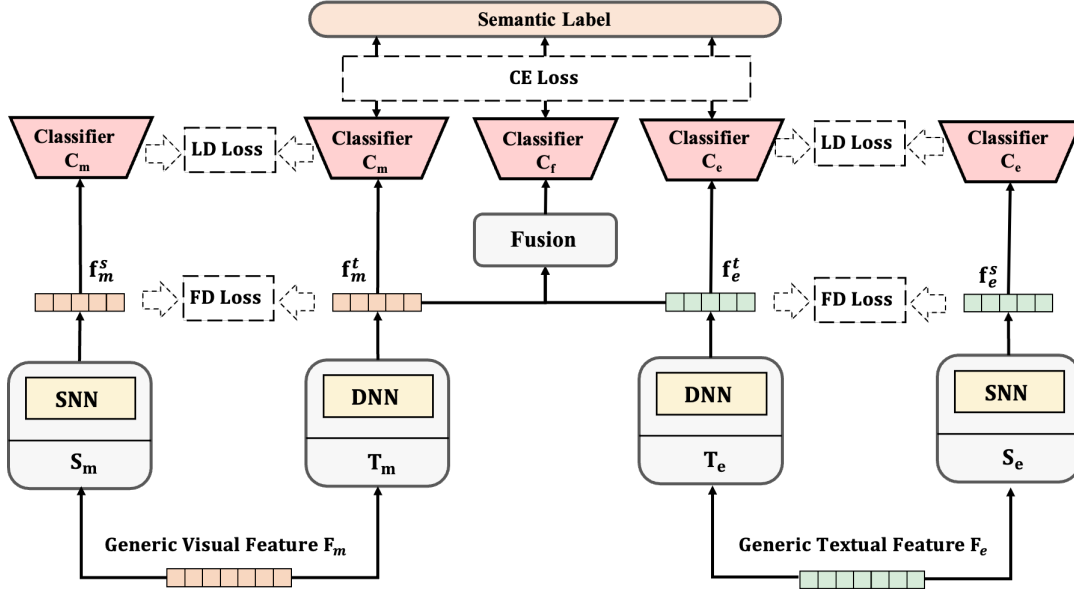


Figure 2: An overview of the proposed Semantic-Guided Feature Distillation (SGFD) method. Taking visual feature extraction as an example, the visual feature extractor in the teacher model extracts features from generic visual features considering both the semantic and complementary information. The Logits-based Distillation (LD) loss and the Feature-based Distillation (FD) loss are employed to transfer the knowledge to the student model.

features from generic multimodal features for recommendation. As shown in Fig. 2, the proposed method consists of a teacher model and a student model. In the teacher model, the DNN-based feature extractors (T_e and T_m) for textual and visual features are designed to extract effective features considering the semantic and complementary information. In our cases, the feature extractor in the student model replaces the original extractor in multimodal recommendation models. In other words, the SNN-based feature extractors (S_e and S_m) in the student model are part of multimodal recommendation models. Therefore, it can obtain the semantic information from teacher extractors by the adopted knowledge distillation technique and benefit from the user behaviors. Next, we detail these two models and the knowledge distillation technique.

3.2.1 Teacher Model. Since semantic information is often overlooked while extracting features from multimodal data for recommendation, we propose a teacher model that extracts effective features considering the semantic information of items. Specifically, we leverage semantic labels (e.g., category labels) as supervised information to classify the extracted features using classifiers. Our proposed method employs three distinct classifiers $C_e(\cdot)$, $C_m(\cdot)$ and $C_f(\cdot)$ for textual, visual, and fused features. Moreover, considering that the generic modality feature is highly complex and high dimensional, we utilize deep neural networks as feature extractors for visual and textual features, denoted as $T_m(\cdot)$ and $T_e(\cdot)$, respectively.

Extracting Semantic-aware Features. Take the generic visual feature $F_m \in \mathcal{R}^{D_m}$ as an example, we first feed F_m into the feature extractor to obtain valuable features $f_m^t \in \mathcal{R}^{d_m}$:

$$f_m^t = T_m(F_m), \quad (2)$$

where T_m is the feature extractor that consists of a deep neural network with k layers³. Similarly, we obtain the textual feature f_e^t from the generic textual feature using the textual feature extractor T_e . The extracted visual and textual features are then separately fed into the classifier $C_m(\cdot)$ and $C_e(\cdot)$. Take the visual feature f_m^t as an example, it is formulated as follows:

$$P_m^t = C_m(f_m^t; \tau) = \text{softmax}((W_m f_m^t + b_m)/\tau), \quad (3)$$

where P_m^t represents the probability distribution output for the visual feature. W_m and b_m are the trainable weight matrices and bias vectors, respectively. It should be noted that the classifier parameters for each modality are trained independently. Here, τ refers to a temperature hyperparameter that controls the smoothness of the distribution. Similarly, the probability distribution output for the textual feature P_e^t can be obtained.

The feature extractors are optimized using cross-entropy (CE) loss. The combined CE loss for textual and visual features is formulated as follows:

$$\mathcal{L}_{TCE} = (\text{CE}(P_m^t, y) + \text{CE}(P_e^t, y))/2, \quad (4)$$

where $\text{CE}(\hat{y}, y) = -\sum_{i=1}^C y_i \cdot \log(\hat{y}_i)$ denotes the cross-entropy loss. C is the number of classes, y is the ground-truth label, and \hat{y} is the predicted probability distribution.

Exploiting Complementary Information. To enhance the ability of teacher extractors to extract effective features of each

³ k is calculated by $k = \log_{\delta} \frac{D_m}{d_m}$, where δ is a hyperparameter that controls the number of layers based on the input dimension of each modality.

modality, we first leverage complementary information between multiple modalities by fusing textual and visual features. As shown in Fig. 2, taking the output of each modality extractor as input, we fuse the vectors as follows:

$$\mathbf{f}_u = \sigma(\mathbf{W}_u[\mathbf{f}_m^t; \mathbf{f}_e^t]) + \mathbf{b}_u, \quad (5)$$

where \mathbf{W}_u and \mathbf{b}_u denote the trainable weight matrix and bias vector. $\sigma(\cdot)$ is the activation function and ReLU is adopted because of its biologically plausible and non-saturated property [46]. Then, the fused feature \mathbf{f}_u is fed into the classifier $C_f(\cdot)$, and the CE loss function of the fused feature is defined as follows:

$$\mathbf{P}_u = C_f(\mathbf{f}_u; \tau), \quad (6)$$

$$\mathcal{L}_{FCE} = \text{CE}(\mathbf{P}_u, \mathbf{y}). \quad (7)$$

The weight matrix and the bias vector of $C_f(\cdot)$ are trainable and specifically trained with fused features to capture the complementary information between multiple modalities. The temperature hyperparameter τ is adopted similarly as they are in Eq. 4.

3.2.2 Student Model. To reduce the negative effect of the data sparsity problem on feature extraction in recommendation, we utilize shallow neural networks (SNNs) as feature extractors for generic multimodal features. The feature extractors for visual and textual features in the student model are denoted as $S_m(\cdot)$ and $S_e(\cdot)$, respectively. For the generic visual feature \mathbf{F}_m , the recommendation-oriented visual feature \mathbf{f}_m^s is obtained by:

$$\mathbf{f}_m^s = S_m(\mathbf{F}_m). \quad (8)$$

The textual feature \mathbf{f}_e^s can be obtained in the same manner. Then, the visual and textual features $\mathbf{f}_m^s, \mathbf{f}_e^s$ are fed into the classifier $C_m(\cdot)$ and $C_e(\cdot)$ respectively. Taking the visual feature as an example, it is formulated as:

$$\mathbf{P}_m^s = C_m(\mathbf{f}_m^s; \tau). \quad (9)$$

By following this way, we can obtain \mathbf{P}_e^s for the textual feature. Note that the classifiers adopted are the same as in Eq. 3.

3.2.3 Knowledge Distillation. The proposed teacher model, which utilizes deep neural networks, is capable of extracting semantic-aware features from the generic features of each modality. On the other hand, the student model, employing SNNs, can alleviate the negative effect of the data sparsity problem on feature extraction in recommendation. To leverage the strengths of both the teacher and student models, we employ the knowledge distillation technique to transfer the response- and feature-based knowledge encoded in the teacher model to the student model. Next, we detail the two distillation losses used for transferring this knowledge.

Response-Based Knowledge. Response-based knowledge refers to the neural response (logits) generated by the final output layer of the teacher model. In our method, we consider the output of the classifiers as response-based knowledge. Hence, we employ Logits-based distillation (LD) loss to enable the feature extractors in the student model to mimic the predicted label distribution of the feature extractors in the teacher model. The distillation loss between the feature extractors of visual and textual features for both the teacher and student models can be described as follows:

$$\mathcal{L}_{LD} = \|\mathbf{P}_m^t - \mathbf{P}_m^s\|_2^2 + \|\mathbf{P}_e^t - \mathbf{P}_e^s\|_2^2. \quad (10)$$

Table 1: Basic statistics of the experimental datasets.

Dataset	#user	#item	#label	#interactions	sparsity
Office	4,874	7,279	54	52,957	99.85%
Clothing	18,209	17,318	26	150,889	99.98%
Toys Games	18,748	11,672	19	161,653	99.97%

Note that we do not force the student model to match the prediction distribution of the teacher model; instead, we expect the student model to gain semantic information via knowledge transfer. Thus, following the previous work [15], we use soft targets to embody the informative knowledge derived from the teacher model.

Feature-Based Knowledge. Given the capacity of deep neural networks for feature extraction, the output from intermediate layers can serve as knowledge to guide the training of the student model. In our method, we employ the Feature-based Distillation (FD) loss to transfer the visual and textual feature-based knowledge from the teacher model to the student model. The distillation loss can be formulated as follows:

$$\mathcal{L}_{FD} = \|\mathbf{f}_m^t - \mathbf{f}_m^s\|_2^2 + \|\mathbf{f}_e^t - \mathbf{f}_e^s\|_2^2. \quad (11)$$

3.3 Optimization

In this work, we target the top- n recommendation and equip our proposed SGFD with three multimodal recommendation models (MAML [22], GRCN [44], BM3 [51]). However, these three recommendation models have different objective functions. Therefore, we refer to the loss of objective functions of these three recommendation models as \mathcal{L}_{REC} .

For our proposed SGFD, the objective function consists of the Cross-Entropy Loss in the teacher model and the Knowledge Distillation Loss. Specifically, the total Cross-Entropy Loss is formulated as:

$$\mathcal{L}_{CE} = \mathcal{L}_{TCE} + \mathcal{L}_{FCE}. \quad (12)$$

The Knowledge Distillation Loss utilized for knowledge transfer is formulated as:

$$\mathcal{L}_{KD} = \mathcal{L}_{LD} + \mathcal{L}_{FD}. \quad (13)$$

The student model of the proposed SGFD is jointly trained with the multimodal recommendation model in an end-to-end manner. Overall, the final loss function of the recommendation model equipped with SGFD is,

$$\mathcal{L}_{SGFD} = \mathcal{L}_{REC} + \lambda_1 \mathcal{L}_{KD} + \lambda_2 \mathcal{L}_{CE}, \quad (14)$$

where λ_1 and λ_2 are used to adjust the weight of the knowledge distillation loss and cross-entropy loss.

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. In experiments, we utilize the Amazon review dataset⁴ for experimental evaluation as in previous studies [22, 26]. This public dataset contains user interactions (review, rating, helpfulness votes, etc.) on items and the item metadata (descriptions,

⁴<http://jmcauley.ucsd.edu/data/amazon>.

attributes, images, etc.) on 24 product categories. To verify the effectiveness of our proposed method, we choose the three per-category datasets (*Office*, *Clothing* and *Toys&Games*) for performance evaluation under different sizes and sparsity levels. In these datasets, we treat each rating as a positive user interaction. In contrast, we treat all the items the user did not interact with as negative. And then, we pre-processed the dataset with a 5-core setting on both items and users in our settings. All user reviews and item images are used as item-side information. Table 1 shows the basic statistics of the four datasets.

In this work, we focus on recommending a set of top- N ranked items that will appeal to the user. Specifically, we randomly sampled 80% of the interactions of each user to create the training set and retained 20% of the user interactions for testing.

4.1.2 Baselines. To demonstrate the effectiveness of our proposed SGFD, we first equipped it on three existing multimodal recommendation models: MAML [22], GRCN [44] and BM3 [51]. Then, we compared the performance of newly built models (SGFD_{MAML}, SGFD_{GRCN}, SGFD_{BM3}) with the following baselines, including general CF recommendation models (e.g., NeuMF [14], CML [16], NGCF [40] and DGCF [41]) and multimodal recommendation models (e.g., MAML, GRCN, BM3).

4.1.3 Evaluation Metrics. Two widely-used evaluation metrics are adopted for top- n recommendation: *Recall* and *Normalized Discounted Cumulative Gain* (NDCG) [12]. Recommendation accuracy is calculated for each metric based on the top 20 results. Note that the reported results are the average values across all testing users.

4.1.4 Parameter Settings. The PyTorch framework⁵ is adopted to implement the proposed model. In our experiments, all hyperparameters are carefully tuned. For all datasets, the embedding size of the users and items is set to 128. The mini-batch size is fixed to 1024. The learning rate for the optimizer is searched from $\{10^{-5}, 10^{-4}, \dots, 10^{-1}\}$, and the model weight decay is searched in the range $\{10^{-5}, 10^{-4}, \dots, 10^{-1}\}$. For knowledge transfer, the temperature τ is tuned from $\{5, 10, 20, 50, 100\}$, while for classification, it is set to 1. The layer number of the neural networks in the teacher model δ is tuned from $\{2, 4, 6, 8\}$. Moreover, the weights λ_1 and λ_2 are searched in the range $\{10^{-3}, 10^{-2}, \dots, 10^{-1}\}$. Besides, the early stopping strategy is adopted. Specifically, the training process will stop if recall@20 does not increase for 20 successive epochs.

4.2 Study of Multimodal Feature Extraction in Recommendation

In this section, we primarily investigate the influence of different feature extractors for multimodal features on recommendation performance. We compare the performance of four feature extraction methods with our proposed SGFD based on GRCN, and the results are reported in Fig. 3. Our analyses are based on performance comparisons to the following methods:

- **GRCN₁** [44]: It is the vanilla GRCN method that adopts the shallow neural network as the feature extractor.
- **GRCN_m**: The vanilla GRCN method adopts the deep neural network as the feature extractor.

⁵<https://pytorch.org>.

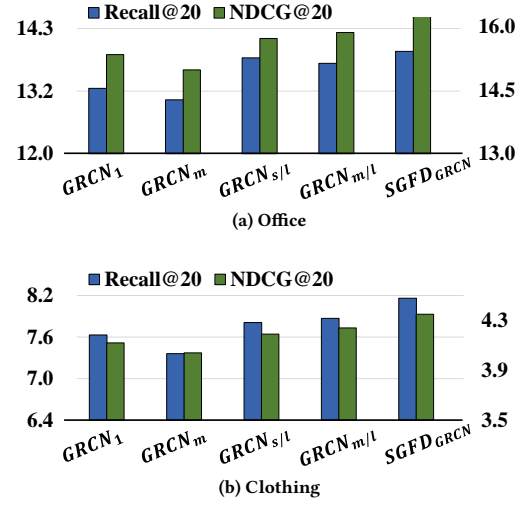


Figure 3: Performance comparison of GRCN equipped with different feature extraction methods. Notice that the values are reported by percentage with '%' omitted.

- **GRCN_{s/l}**: Given the vanilla GRCN method, we employ the teacher model of SGFD with shallow neural networks as the feature extractor in GRCN.
- **GRCN_{m/l}**: Given the vanilla GRCN method, we employ the teacher model of SGFD with the deep neural networks as the feature extractor in GRCN.

From the results, we can see that GRCN₁ yields better performance than GRCN_m on *Office* and *Clothing*. Furthermore, the performance degradation of GRCN_m in the *Office* is less compared to *Clothing*. This can be attributed to the fact that the feature extractor in most multimodal recommendation models is jointly optimized with recommendation models. Consequently, the feature extractor will suffer the data sparsity problem when the user-item interactions are sparsity. As *Office* is denser, the data sparsity problem less impacts the feature extraction model. The variant GRCN_{s/l} outperforms GRCN₁ and GRCN_m across all the datasets, which indicates that the semantic information has positive advancement in extracting valuable features for recommendation. More importantly, by adopting multiple neural networks and introducing the complementary information of multimodal features, the variant GRCN_{m/l} has an explainable improvement on *Clothing*. This demonstrates that the complementary information is facilitated to reduce redundant information derived from the feature extractor in each modality. But, it still underperforms GRCN_{s/l} on *Office* in terms of Recall. The reason might be that it also suffers from the data sparsity problem when jointly training the extractor with recommendation models. The SGFD_{GRCN} outperforms all variants over all datasets, which demonstrates the effectiveness of our proposed SGFD.

4.3 Performance Comparison

We report the performance of three multimodal recommendation models equipped with our proposed SGFD and all the competitors over the three datasets in table 2. Our proposed approach achieves

Table 2: Performance of our SGFD model and the competitors over three datasets. Notice that the values are reported by percentage with '%' omitted.

Datasets	Office		Clothing		Toys&Games	
Metrics	Recall	NDCG	Recall	NDCG	Recall	NDCG
NeuMF	6.05	3.37	1.89	0.80	2.53	1.28
CML	12.29	13.95	4.09	2.24	12.27	11.60
NGCF	8.14	4.79	4.66	2.16	9.70	5.87
DGCF	12.06	11.05	7.45	4.05	12.62	10.85
MAML	13.33	14.12	6.89	3.70	12.84	12.09
SGFD _{MAML}	14.22*	14.87*	7.18*	3.89*	13.23*	12.60*
Improv.	6.68%	5.31%	4.21%	5.14%	3.04%	4.22%
GRCN	13.20	15.33	7.63	4.12	11.85	11.23
SGFD _{GRCN}	13.88*	16.26*	8.16*	4.35*	12.59*	11.93*
Improv.	5.15%	6.07%	6.95%	5.58%	6.24%	6.23%
BM3	13.67	15.36	7.43	4.21	12.22	11.92
SGFD _{BM3}	14.07*	15.86*	7.84*	4.57*	12.54*	12.94*
Improv.	2.93%	3.25%	5.52%	8.56%	2.62%	8.56%

The symbol * denotes that the improvement is significant with $p - value < 0.05$ based on a two-tailed paired t-test.

significant performance improvements over all the competitors across all the datasets in terms of different metrics. From the experiment results, we achieve some compelling observations.

Among all competitors, NeuMF, CML, NGCF and DGCF are solely trained using user-item interactions. NeuMF achieves promising performance in comparison to traditional MF-based methods [14], due to the capability of deep neural networks to model the non-linear interactions between users and items. The dot product cannot well model fine-grained user preferences, as it does not obey the triangle inequality. Therefore, CML replaces the dot product with metric distance, improving performance compared to NeuMF. Benefiting from the high-order information, NGCF accomplishes state-of-the-art results. Moreover, DGCF achieves better performance than NGCF by leveraging the disentangled representation technique to learn robust and independent user and item embeddings. MAML, GRCN, and BM3, which utilize multimodal features, generally achieve better performance than those methods relying solely on user-item interactions. Specifically, MAML outperforms CML by modeling users' diverse preferences by leveraging multimodal features of items. GRCN surpasses NGCF by employing modality features to discover and prune potential false-positive edges on the user-item interaction graph. BM3 consistently performs better than all general CF baselines over all the datasets. It should be credited to the joint optimization of three multimodal objectives to learn the representations of users and items.

To demonstrate the effectiveness of our proposed method, we equip MAML, GRCN and BM3 with SGFD, which are denoted as SGFD_{MAML}, SGFD_{GRCN} and SGFD_{BM3}, by employing the feature extractor in student model to replace the original feature extractor in the recommendation models. Note that our approach focuses on distilling knowledge from the teacher to the student model. Hence, the users and items representation learning approaches of the base models (MAML, GRCN and BM3) are not modified. The performance of SGFD_{MAML}, SGFD_{GRCN} and SGFD_{BM3} are reported in Table 2. From the results, we can see that both SGFD_{MAML},

Table 3: Performance of our proposed SGFD method equipped on GRCN and their variants over three datasets. The best results are highlighted in bold. Notice that the values are reported by percentage with '%' omitted.

Datasets	Office		Clothing		ToysGames	
Metrics	Recall	NDCG	Recall	NDCG	Recall	NDCG
SGFD _l	13.30	15.80	7.69	4.06	12.01	11.31
SGFD _f	13.53	15.71	7.86	4.18	11.55	10.92
SGFD _e	13.71	15.73	7.97	4.28	12.08	11.36
SGFD _s	13.76	16.07	8.02	4.30	12.22	11.50
SGFD	13.88	16.26	8.16	4.35	12.59	11.93

SGFD_{GRCN} and SGFD_{BM3} obtain a significant improvement over all datasets. We credit this to the combined effects of the following three aspects. Firstly, the teacher model can extract semantic-aware features from generic multimodal features. Secondly, the knowledge encoded in the teacher model is transferred to the student model by the response- and feature-based knowledge distillation losses. Thirdly, the feature extractors with shallow neural networks in the student model replace the original feature extractors in the base models. Based on this, existing multimodal models can ease the negative impact of the data sparsity problem on representation learning.

4.4 Ablation Study

In this section, we examine the contribution of different components to the performance of our method, which is equipped with the recommendation method (GRCN). Our analyses are based on performance comparisons with the following variants of our method.

- **SGFD_l**: It is a variant of SGFD that uses only logits-based distillation loss for knowledge distillation.
- **SGFD_f**: It is a variant of SGFD that uses only feature-based distillation loss for knowledge distillation.
- **SGFD_e**: This variant removes the complementary information exploited from multimodal features in the teacher model.
- **SGFD_s**: This variant adopts the shallow neural network as the feature extractor in the teacher model.

From the results, the SGFD_l and SGFD_f gain different improvements over three datasets, which indicates the benefits from the response-based knowledge and the feature-based knowledge are different. Benefiting from using both the logits- and the feature-based distillation loss, SGFD_e achieves a remarkable performance across all the datasets. SGFD_s outperforms SGFD_e across all datasets in terms of NDCG. This is because the complementarity between different modality features is taken into account during the feature extraction process. Our proposed SGFD method achieves superior performance across all datasets. It demonstrates the effectiveness of using the feature extractor with deep neural networks to extract the semantic-aware feature.

4.5 Visualization of Knowledge Transfer

To confirm the transfer of knowledge from the teacher model to the student model, we visualized the review feature and image feature vectors extracted by the original feature extractor in GRCN

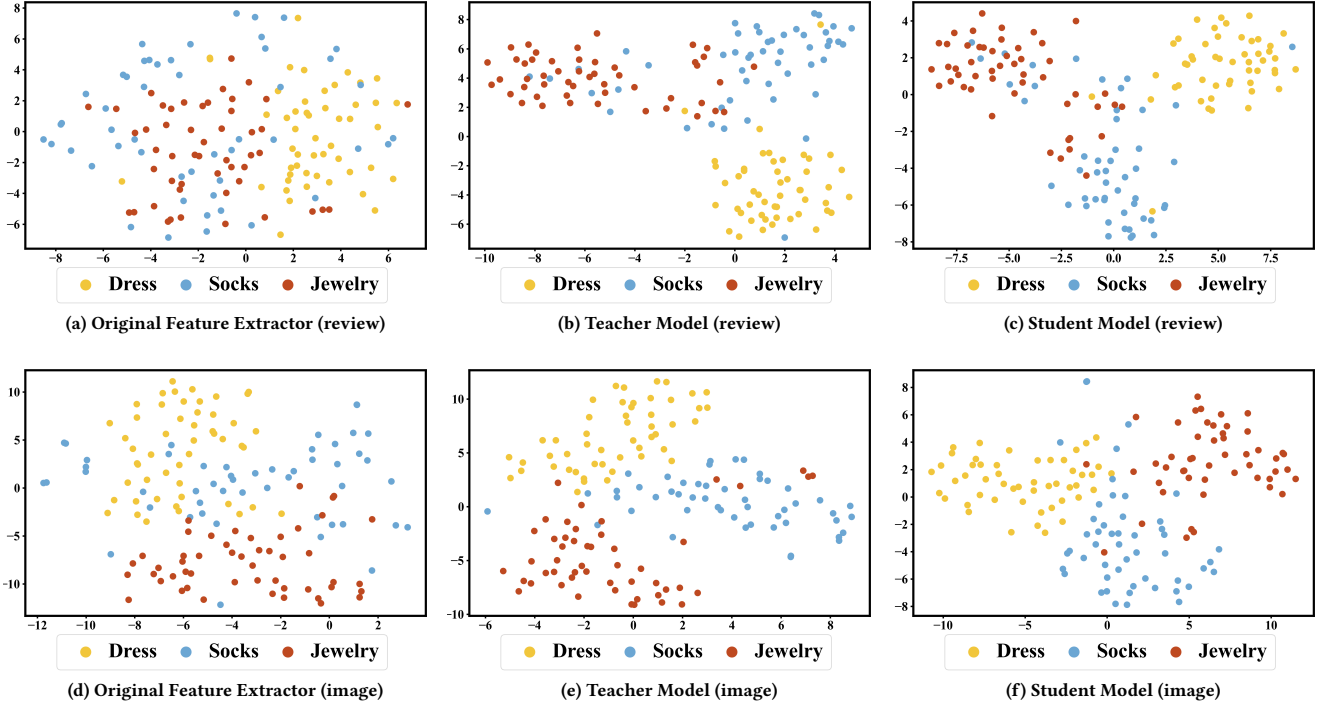


Figure 4: Visualization of the vectors extracted from generic review and textual features. The vectors are derived from the original feature extractor in GRCN, feature extractors in teacher and student models of SGFD_{GRCN}. In addition, the items are sampled from three categories of *Clothing*: Socks (Orange), Dress (Blue) and Jewelry (Green).

as well as the feature extractors in the teacher and student model of SGFD_{GRCN}, respectively.

In Fig. 4, the top three figures show the feature vectors extracted from the generic review feature of items. Meanwhile, the following three figures illustrate the vectors extracted from the generic image feature of items. These items are randomly sampled from three categories of *Clothing* (accessories, body jewelry and socks). Each dot in the figures denotes the extracted vector from the feature of an item. Besides, the dots with the same color in all figures denote the same category of items. The obtained vectors are clustered and visualized via the t-Distributed Stochastic Neighbor Embedding (t-SNE). As shown in Fig. 4, for the feature vectors extracted from different extractors, those items of the same category from the teacher extractor are closer than the original extractor, and the boundaries between categories are also more straightforward in Fig. 4(b). This is attributed to using semantic information to guide the feature extraction process in the teacher model. Besides, the clustering results shown in Fig. 4(c) demonstrate that the knowledge encoded in the teacher model is transferred to the student model. A similar trend can be observed from the clustering result of feature vectors extracted from generic image features based on the three extractors.

5 CONCLUSION

In this work, we propose a novel model-agnostic approach, named Semantic-Guided Feature Distillation (SGFD), which can robustly

extract effective recommendation-oriented features from generic modality features for recommendation. In our approach, feature extractors in the teacher model are proposed to extract effective features considering both the semantic information of items from labels and the complementary information of multiple modalities. Moreover, to reduce the negative effect of the data sparsity problem on feature extraction in recommendation, we employ the student model consisting of feature extractors with shallow neural networks. By transferring the knowledge encoded in the teacher model to the student model and replacing the original feature extractor in recommendation methods with the feature extractors in the student model of our SGFD, effective recommendation-oriented features can be obtained. The experimental results on three real-world datasets demonstrate our approach can significantly improve the recommendation accuracy of multimodal recommendation methods.

6 ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative, the National Natural Science Foundation of China under Grants 62272254, the Shandong Project towards the Integration of Education and Industry under Grants 2022PY009 and 2022PYI001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

- [1] Sungsoo Ahn, Shell Xu Hu, Andreas C. Damianou, Neil D. Lawrence, and Zhenwen Dai. 2019. Variational Information Distillation for Knowledge Transfer. *CVPRs* (2019), 9155–9163.
- [2] Robert M Bell and Yehuda Koren. 2007. Lessons from the netflix prize challenge. *SIGKDD* 9, 2 (2007), 75–79.
- [3] Rose Catherine and William Cohen. 2017. TransNets: Learning to Transform for Recommendation. In *RecSys*. ACM, 288–296.
- [4] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *WWW*. IW3C2, 1583–1592.
- [5] Jingwu Chen, Fuzhen Zhuang, Xin Hong, Xiang Ao, Xing Xie, and Qing He. 2018. Attention-driven Factor Model for Explainable Personalized Recommendation. In *SIGIR*. ACM, 909–912.
- [6] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan Kankanhalli. 2018. A³NCF: An Adaptive Aspect Attention Model for Rating Prediction. In *IJCAI*. AAAI Press, 3748–3754.
- [7] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. 2018. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *WWW*. IW3C2, 639–648.
- [8] Zhiyong Cheng, Sai Han, Fan Liu, Lei Zhu, Zan Gao, and Yuxin Peng. 2023. Multi-Behavior Recommendation with Cascading Graph Convolution Networks. In *WWW*. ACM, 1181–1189.
- [9] Jin Yao Chin, Kaiqi Zhao, Shafiq Joty, and Gao Cong. 2018. ANR: Aspect-based Neural Recommender. In *CIKM*. ACM, 147–156.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*. ACL, 4171–4186.
- [11] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*. AAAI Press, 144–150.
- [12] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. TriRank: Review aware Explainable Recommendation by Modeling Aspects. In *CIKM*. ACM, 1661–1670.
- [13] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR*. ACM, 639–648.
- [14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. IW3C2, 173–182.
- [15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *stat* 1050 (2015), 9.
- [16] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. 2017. Collaborative metric learning. In *WWW*. IW3C2, 193–201.
- [17] Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. 2013. Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *ICMR*. ACM, 105–112.
- [18] SeongKu Kang, Junyoung Hwang, Wonbin Kweon, and Hwanjo Yu. 2020. DE-RRD: A Knowledge Distillation Framework for Recommender System. In *CIKM*. ACM, 605–614.
- [19] Nikos Komodakis and Sergey Zagoruyko. 2017. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *ICLR*. 1–13.
- [20] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [21] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. 2022. Disentangled Multimodal Representation Learning for Recommendation. *TMM* (2022), 1–11.
- [22] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User Diverse Preference Modeling by Multimodal Attentive Metric Learning. In *MM*. ACM, 1526–1534.
- [23] Fan Liu, Zhiyong Cheng, Lei Zhu, Zan Gao, and Liqiang Nie. 2021. Interest-Aware Message-Passing GCN for Recommendation. In *WWW*. ACM, 1296–1305.
- [24] Fan Liu, Zhiyong Cheng, Lei Zhu, Chenghao Liu, and Liqiang Nie. 2022. An Attribute-Aware Attentive GCN Model for Attribute Missing in Recommendation. *TKDE* 34, 9 (2022), 4077–4088.
- [25] Han Liu, Yinwei Wei, Jianhua Yin, and Liqiang Nie. 2023. HS-GCN: Hamming Spatial Graph Convolutional Networks for Recommendation. *TKDE* 35, 6 (2023), 5977–5990.
- [26] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*. ACM, 165–172.
- [27] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*. ACM, 43–52.
- [28] Pranay Mukherjee, Abhirup Das, Ayan Kumar Bhunia, and Partha Pratim Roy. 2019. Cogni-Net: Cognitive Feature Learning Through Deep Visual Perception. In *ICIP*. IEEE, 4539–4543.
- [29] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *ICLR*. 1–13.
- [30] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*. 1–14.
- [31] Teng Sun, Chun Wang, Xuemeng Song, Fuli Feng, and Liqiang Nie. 2022. Response Generation by Jointly Modeling Personalized Linguistic Styles and Emotions. *TOMM* 18, 2 (2022), 52:1–52:20.
- [32] Teng Sun, Wenjie Wang, Liqiang Nie, Yiran Cui, Xuemeng Song, and Liqiang Nie. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *MM*. ACM, 15–23.
- [33] Yunzhi Tan, Min Zhang, Yiqun Liu, and Shaoping Ma. 2016. Rating-boosted latent topics: Understanding users and items with ratings and reviews. In *IJCAI*. AAAI Press, 2640–2646.
- [34] Jiaxi Tang and Ke Wang. 2018. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *SIGKDD*. ACM, 2289–2298.
- [35] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Representation Distillation. In *ICLR*. 1–19.
- [36] Shaokun Wang, Tian Gan, Yuan Liu, Jianlong Wu, Yuan Cheng, and Liqiang Nie. 2023. Micro-Influencer Recommendation by Multi-Perspective Account Representation Learning. *TMM* 25 (2023), 2749–2760.
- [37] Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. 2017. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *WWW*. IW3C2, 391–400.
- [38] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. 2022. User-controllable recommendation against filter bubbles. In *SIGIR*. ACM, 1251–1261.
- [39] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative recommendation: Towards next-generation recommender paradigm. *arXiv preprint arXiv:2304.03516* (2023).
- [40] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *SIGIR*. ACM, 165–174.
- [41] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled Graph Collaborative Filtering. In *SIGIR*. ACM, 1001–1010.
- [42] Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized Hashtag Recommendation for Micro-videos. In *MM*. ACM, 1446–1454.
- [43] Yinwei Wei, Wenqi Liu, Fan Liu, Xiang Wang, Liqiang Nie, and Tat-Seng Chua. 2023. LightGT: A Light Graph Transformer for Multimedia Recommendation. In *SIGIR*. ACM, 1508–1517.
- [44] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *MM*. ACM, 3541–3549.
- [45] Jae woong Lee, Minjin Choi, Jongwuk Lee, and Hyunjung Shim. 2019. Collaborative Distillation for Top-N Recommendation. *ICDM* 2019 (2019), 369–378.
- [46] Glorot Xavier and Bengio Yoshua. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*. JMLR, 249–256.
- [47] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *SIGKDD*. ACM, 353–362.
- [48] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *MM*. ACM, 3872–3880.
- [49] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *CIKM*. ACM, 1449–1458.
- [50] Xinxiao Zhao, Fan Liu, Hao Liu, Mingzhu Xu, Haoyu Tang, Xueqing Li, and Yupeng Hu. 2023. CoGCN: co-occurring item-aware GCN for recommendation. *NCAA* (2023), 1–14.
- [51] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap Latent Representations for Multimodal Recommendation. In *WWW*. ACM, 845–854.