Information Sciences 653 (2024) 119760

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Self-supervised Multimodal Graph Convolutional Network for collaborative filtering

Sungjune Kim [a], Seongjun Yun [c], Jongwuk Lee [b], Gyusam Chang [a], Wonseok Roh [a], Dae-Neung Sohn [d], Jung-Tae Lee [d], Hogun Park [b,*], Sangpil Kim [a,*]

[a] Department of Artificial Intelligence, Korea University, South Korea
[b] Department of Artificial Intelligence, Sungkyunkwan University, South Korea
[c] Department of Computer Science and Engineering, Korea University, South Korea
[d] Naver Corp., South Korea

ARTICLE INFO

Keywords:
Multimodal data
Self-supervised learning
Graph neural networks

ABSTRACT

Collaborative filtering (CF) is a central solution for capturing various user-item relationships in building recommender systems. However, when the relationships are sparsely observed, it is challenging to obtain enough signals to infer precise user preferences. Recent studies have attempted to address the sparsity issue by incorporating multimodal information (e.g., image and text) into CF models. However, existing methods mainly focus on capturing modal-specific user preference with multiple unimodal graphs, ignoring the complex nature of user behavior, which is determined by an intricate fusion of multimodal information. Therefore, we develop a Self-supervised Multimodal Graph Convolutional Network (SMGCN), which aims to learn the cross-modal user preferences over multiple modalities with an expressive multimodal fusion on a single graph.

More importantly, to facilitate and enhance multimodal fusion in SMGCN, we devise two novel self-supervised learning techniques. 1) Collaborative Multimodal Alignment (CMA) uses contrastive learning to align the domain-specific multimodal semantics with the user-item relational semantics. 2) Multimodal Consistency Regularization (MCR) alleviates the sensitivity on a certain modality and increases model robustness. The experimental results demonstrate that our model consistently outperforms advanced multimodal models on three benchmark datasets.

1. Introduction

Recommender systems support personalized user experiences in online streaming services, e-commerce, and social media [45, 21]. Collaborative filtering (CF) is crucial for developing modern recommendation systems [27,16]. The primary concept of CF is to capture the *homophily of users/items*, indicating that users with similar user-item interactions are likely to share similar user preferences. The main advantage of CF models is that they utilize only past user-item interactions to estimate hidden user preferences. Therefore, capturing collaborative signals with additional information about users/items is beneficial.

In conventional methods, matrix factorization (MF) models [25] decompose a user-item interaction matrix as the inner product of the latent user and item embedding vectors. In neural collaborative filtering [12], this concept is further developed by applying

* Corresponding authors.
E-mail addresses: hogunpark@skku.edu (H. Park), spk7@korea.ac.kr (S. Kim).

https://doi.org/10.1016/j.ins.2023.119760
Received 31 March 2023; Received in revised form 7 September 2023; Accepted 1 October 2023

Available online 5 October 2023

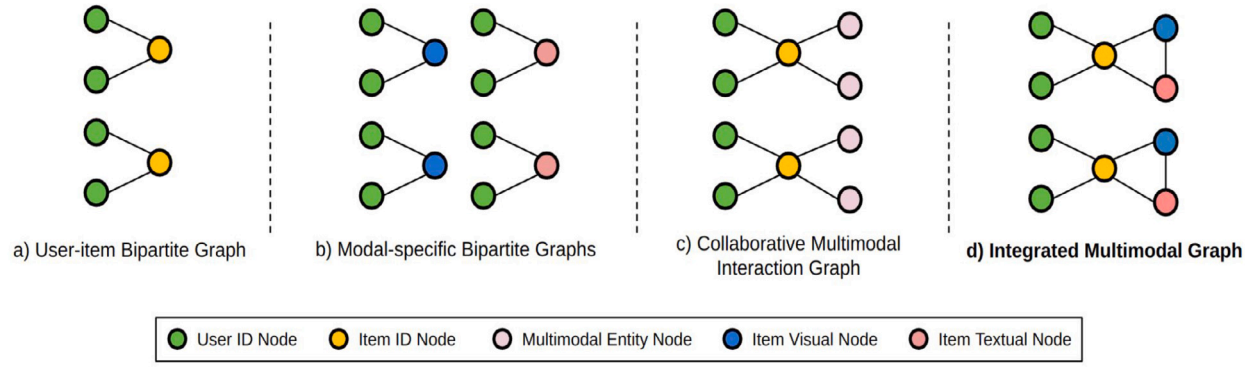


Fig. 1. An illustration of the structural novelty of Integrated Multimodal Graph (IMGraph). IMGraph is novel in that it enables an explicit fusion of multiple modalities and collaborative embedding representations. Best viewed in color.

neural networks for a more complex understanding of interactions. Considering the recent success of graph neural networks (GNNs) in diverse domains [38,33,32,42], CFs have also achieved significant progress by reformulating user-item interaction as a bipartite graph to aggregate collaborative signals from neighboring nodes [34,13,35].

Despite its effectiveness, the high dependency of CFs on the existence of historical user-item interaction makes them inevitably face the data sparsity problem. Fortunately, owing to the heterogeneity of data on the web, recent works have been incorporating multimodal information (e.g. image and text) into the CF frameworks. Early studies [11,3] extend MF models to exploit multimodal features. Recently, graph-based multimodal CF methods show promising results, where GNNs are applied to capture modal-specific user preferences by constructing a user-item bipartite graph for each modality [30,37,29,40,36].

However, existing methods have several limitations. First, they overlook the complex multimodal nature of user behavior. For example, users are generally influenced by both image and text of a certain item. However, MMGCN [37] and the following studies [30,29,40,36] designed separate graph for each modality. Therefore, these methods do not consider how multiple modalities jointly affect user preferences. Furthermore, when the user-item interaction is sparsely observed, it is challenging to acquire modal-specific user preferences because these methods simply replace the ID embeddings with multimodal features. Considering that embedding updates still depend on the existence of interactions, they may not capture modality-relevant collaborative signals.

Motivated by the limitations of the previous works, we develop Self-supervised Multimodal Graph Convolutional Network (SMGCN). The core idea of SMGCN is to learn the cross-modal user preferences over multiple modalities. SMGCN achieves this by constructing a novel Integrated Multimodal Graph (IMGraph), which enables an explicit data fusion of multiple modalities as well as collaborative embedding representations. Our IMGraph is a combined structure of the user-item bipartite graph and item-multimodal graphs where item nodes act as an anchor between two graphs. An item-multimodal graph is a complete graph consisting of an item ID node and its multimodal feature nodes.

Since multimodal features are obtained from the pre-trained encoders of different domains, incorporating the extracted features into the user-item collaborative framework in a supervised manner can be further refined with supportive supervision. To intensify the data fusion of multimodal features with user-item collaborative representations in IMGraph, we develop two novel fusion-oriented self-supervised learning methods.

First, Collaborative Multimodal Alignment (CMA) aims to distill the structural user-item collaborative signals into multimodal representations. By learning to contrast the multimodal features based on the user-item interaction history, the model achieves feature alignment between different domains and facilitates multimodal fusion on the graph. Second, Multimodal Consistency Regularization (MCR) is a simple yet effective method that alleviates the potential variance problems caused by the diversity of multimodal features. We utilize the heterogeneity in IMGraph to construct multimodal graph augmentations, and minimize the differences between the output distributions of multiple augmentations. This reduces the dependency on certain features or modalities which increases the model robustness in fusing multimodal data.

In Table 1, we emphasize the novelty and difference of SMGCN over several baselines in various aspects. In the following sections, we explain our SMGCN in detail and report experimental results, showing that SMGCN outperforms state-of-the-art top-K recommender system baselines. Further analysis demonstrates how each component of SMGCN affects and enhances the overall recommendation performance.

In a nutshell, the contributions of this work are summarized as follows:

- We propose a Self-supervised Multimodal Graph Convolutional Network (SMGCN), which addresses the importance of capturing cross-modal user preferences with an expressive multimodal fusion on a single graph.
- We introduce two essential self-supervisory solutions to enhance the multimodal fusion on graphs for collaborative filtering: Collaborative Multimodal Alignment (CMA) and Multimodal Consistency Regularization (MCR).
- Through extensive experiments, we demonstrate that our SMGCN outperforms state-of-the-art baselines in top-K recommendation tasks in three real-world datasets.

Table 1
Comparison of SMGCN with several baselines across different aspects. The symbol * indicates that the graph is modal-specific.

Method	Multimodal	Graph type	Self-supervised learning	
			Contrastive Learning	Consistency Regularization
LightGCN [13]	-	bipartite	-	-
SGL [39]	-	bipartite	✓	-
MMGCN [37]	✓	bipartite*	-	-
MMGCL [40]	✓	bipartite*	✓	-
SLMRec [30]	✓	bipartite*	✓	-
MEGCF [19]	✓	multimodal	-	-
MMSSL [36]	✓	bipartite*	✓	-
SMGCN	✓	multimodal	✓	✓

2. Related work

2.1. GCN-based recommender systems

Graph convolutional networks (GCNs) [38] are extensions of Convolutional Neural Networks (CNNs) [17] to relational data, with applications in diverse domains [18,7,14]. Specifically, recommender systems have largely benefited from GCNs by learning latent user-item relations on graphs [1,34,13,6]. GC-MC [1] proposes a graph convolution-based auto-encoder framework for user-item matrix completion. NGCF [34] adopts the message-passing algorithm to propagate high-order collaborative signals in user-item bipartite graphs. LightGCN [13] further simplifies and improves NGCF by removing the redundant non-linear activation function and linear transformation of the ID embeddings. In MB-CGCN [6], GCN blocks are utilized to explicitly model the multi-behaviors for embedding learning. Despite their effectiveness, GCN-based recommender systems can be further enhanced by incorporating multimodal item information for fine-grained user preference learning.

2.2. Multimodal recommender systems

To alleviate the data sparsity problem of conventional CF models, many efforts have been made to utilize the multimodal information of the items. For example, VBPR [11] incorporates visual features into matrix factorization. Subsequent studies such as ACF [3] and VECF [5] introduce the attention mechanism in CFs to address the item- and component-level implicit feedback. CKE [43] leverages multimodal item features to construct knowledge base embeddings. MKGAT [28] incorporates multimodal features as knowledge graph entities. MMGCN [37] and MGAT [29] adopt GCN techniques and construct a modal-specific bipartite graph to capture user preferences on each modality. In PMGT [20], the authors introduce a deep pre-training method to exploit item multimodal features through unsupervised learning. Similar to our proposed graph structure, MEGCF [19] designs a collaborative multimodal interaction graph by fusing a user-item bipartite graph with an item-entity bipartite graph. However, they did not consider the direct feature interaction between the multimodal nodes, which shows crucial for multimodal fusion in our experimental result (Sec 6.1). The structural differences between IMGraph with previous methods are illustrated in Fig. 1.

2.3. Self-supervised learning for recommender systems

Another efficient strategy to solve the data sparsity problem in recommender systems is to explore complementary supervision through self-supervised learning (SSL) [41]. SGL [39] first introduces the paradigm in the graph-based CF framework by contrasting the views of several graph augmentations. In LightGCL [2], the authors propose a robust and lightweight graph contrastive augmentation with singular value decomposition. Recently, many works integrate the SSL paradigm into multimodal recommendation [30,44,40,36]. MICRO [44] leverages the contrastive framework on multimodal features to discover latent relationships among items. MMGCL [40] and SLMRec [30] enhance the multimodal representation learning by contrasting the augmentations on modal-specific bipartite graphs. MMSSL [36] proposes a modality-aware adversarial self-augmentation technique to effectively learn modality-aware user preferences. However, our method finds multimodal contrastive pairs within our proposed integrated multimodal graph structure to encode rich collaborative signals into the multimodal representations. Furthermore, we exploit additional SSL to regularize the multimodal consistency, which has never been explored in the literature.

3. Background

3.1. Integrated Multimodal Graph (IMGraph)

We first reformulate the historical interaction data as a user-item bipartite graph $\mathcal{G}_B = \{(u,i) \mid u \in \mathcal{U},\ i \in \mathcal{I}\}$, where $\mathcal{U}$ and $\mathcal{I}$ denote the user and item set respectively. Then, we randomly initialize the initial embedding vectors for each user and item as $e_u^{(0)} \in \mathbb{R}^d$ and $e_i^{(0)} \in \mathbb{R}^d$, where d is the embedding dimension. In addition to $\mathcal{G}_B$, we construct *Integrated Multimodal Graph* $\mathcal{G}_\mathcal{M} = \mathcal{G}_B \cup \{(i,v),(i,t),(v,t) \mid i \in \mathcal{I},\ v \in \mathcal{V},\ t \in \mathcal{T}\}$ by incorporating visual and textual features of each item as nodes into the interaction graph

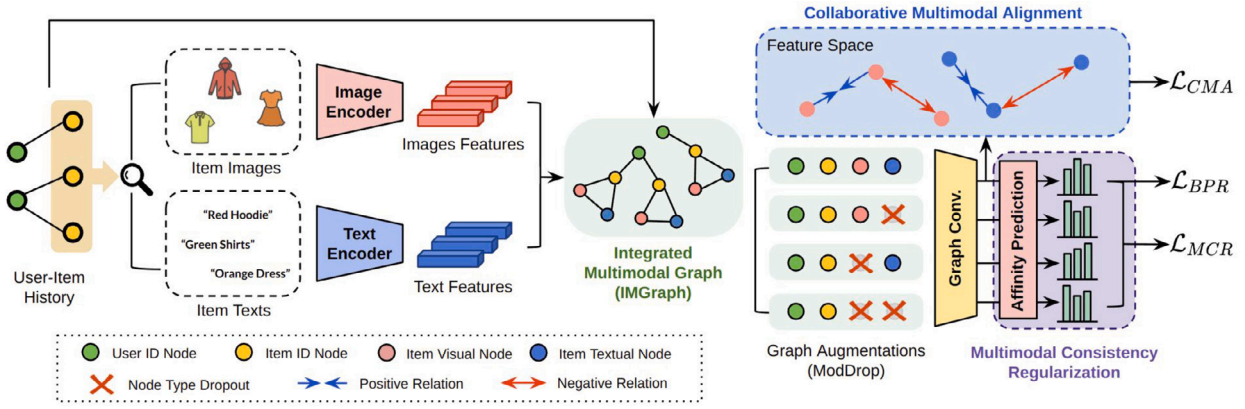


Fig. 2. An illustration of the overall architecture of SMGCN. First we construct an Integrated Multimodal Graph (IMGraph) with multimodal features extracted from each pre-trained encoder. Then, multimodal fusion is achieved through graph convolution and enhanced with two novel self-supervised learning techniques: Collaborative Multimodal Alignment (CMA) and Multimodal Consistency Regularization (MCR). Best viewed in color.

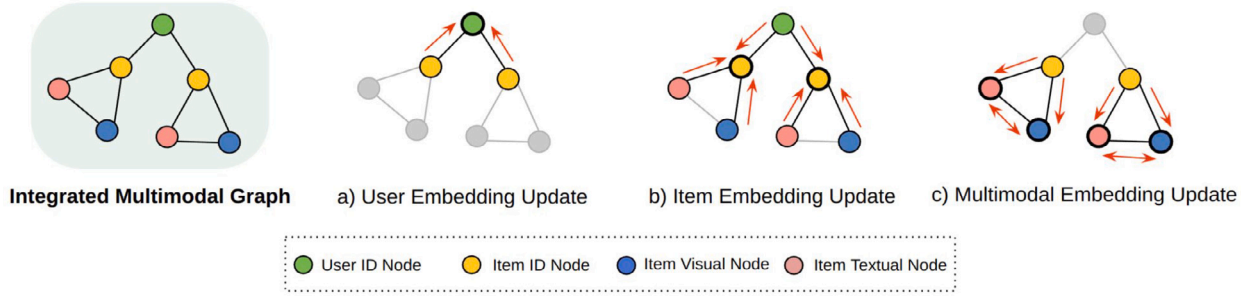


Fig. 3. An illustration of graph convolution over IMGraph. Our method is novel in that multimodal features are treated as nodes in a single graph and it enables direct message passing between different modalities. Best viewed in color.

$\mathcal{G}_{\mathcal{B}}$, where $\mathcal{V}$ and $\mathcal{T}$ denotes the set of visual and textual entities of items. For each item $i \in \mathcal{I}$, the initial multimodal embeddings $e_{i,v}^{(0)}$ and $e_{i,t}^{(0)}$ are extracted from each pre-trained vision and language model (ViT [8] and Sentence-BERT [24]).

3.2. Supervised recommendation tasks

Given an Integrated Multimodal Graph $\mathcal{G}_{\mathcal{M}}$, the supervised task of recommendation is to learn a prediction function $\hat{y}_{ui} = \mathcal{F}(\mathcal{G}_{\mathcal{M}}, E; \Theta)$ which indicates the potential probability that a user u might interact with an item i, where E is the matrix containing all the embedding vectors of the nodes in IMGraph, and Θ denotes the model parameters of the function $\mathcal{F}$. We use this prediction value $\hat{y}_{ui}$ to optimize the model parameters in a supervised manner. Following the convention in recommender systems, we use a pairwise Bayesian Personalized Ranking (BPR) loss [25] to encourage the prediction of an observed interaction to be ranked higher than the unobserved sample.

$$\mathcal{L}_{BPR} = \sum_{(u,i,k)\in\mathcal{O}} -\log \sigma(\hat{y}_{ui} - \hat{y}_{uk}), \tag{1}$$

where $\mathcal{O} = \{(u,i,k)|(u,i) \in \mathcal{O}^+, (u,k) \in \mathcal{O}^-\}$ is the training data with $\mathcal{O}^+$ as observed interactions, and $\mathcal{O}^-$ as unobserved ones.

4. Proposed model

This section introduces SMGCN, a graph-based recommender system that aims at understanding the comprehensive user preference over multiple modalities with expressive data fusion. The overall architecture is illustrated in Fig. 2.

4.1. Graph convolutions on IMGraph

We first introduce a neural graph encoder that performs layer-wise message passing on IMGraph. Unlike commonly used modal-specific bipartite graphs, IMGraph enables data fusion of multiple modalities with graph convolution on a single graph. As shown in Fig. 3, we formulate the embedding equations for a single layer and further generalize it to multiple consecutive layers.

User & Item ID Node Embedding In the IMGraph, the user ID nodes aggregate messages only from their 1-hop neighboring item nodes. We follow LightGCN [13] for the message passing between user and item, where the embedding function for user u from layer l to $l+1$ can be described as:

$$e_u^{(l+1)} = e_u^{(l)} + \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} e_i^{(l)} \tag{2}$$

where $\mathcal{N}_u$ ($\mathcal{N}_i$) is a set of items (users) that are connected with user u (item i). For the item ID nodes, they first follow the aggregation function of users and capture collaborative signals. They also gather messages from their multimodal feature nodes, which are all combined as follows:

$$e_i^{(l+1)} = e_i^{(l)} + \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_u||\mathcal{N}_i|}} e_u^{(l)} + g\Big(\sum_{m \in \{v,t\}} W_m^{(l)} e_{i,m}^{(l)} \Big). \tag{3}$$

Unlike the randomly initialized user and item ID embeddings, the features of the multimodal nodes contain concrete semantics. Therefore, to extract higher-level features of the multimodal information, we apply a non-linear transformation to the messages from the multimodal nodes, where $W_m^{(l)}$ indicates the trainable weight matrix of modality m in the l^{th} layer and $g(\cdot)$ is a non-linear activation function LeakyReLU [22].

Multimodal Node Embedding We treat the multimodal features of items as distinct nodes in IMGraph. Therefore, the multimodal nodes receive messages from their item ID nodes in the embedding layer. Additionally, we directly enable feature interaction between different modalities of the same item, which is essential for complex multimodal fusion on graphs. The described process is formulated as:

$$e_{i,v}^{(l+1)} = e_{i,v}^{(l)} + g(e_i^{(l)} + e_{i,t}^{(l)}), \tag{4}$$

$$e_{i,t}^{(l+1)} = e_{i,t}^{(l)} + g(e_i^{(l)} + e_{i,v}^{(l)}). \tag{5}$$

Prediction Layer After L successive layers of propagation, we obtain the final features of user and item $e_u^{(L)}$ and $e_i^{(L)}$, which are precisely discriminated by the rich multimodal semantics of the items. The affinity between each user u and item i is calculated by the inner product of the two final representation vectors:

$$\hat{y}_{ui} = e_u^{(L)\intercal} e_i^{(L)}. \tag{6}$$

Here, we follow the mainstream practice of using the inner product as the prediction score for simplicity [34,13,37], and leave more sophisticated prediction functions to be explored in future work.

4.2. Collaborative Multimodal Alignment

Multimodal fusion on IMGraph can be further enhanced by enabling feature alignment between multimodal features and collaborative embeddings. Collaborative Multimodal Alignment (CMA) achieves this by distilling the structural user-item collaborative signals into multimodal representations through contrastive learning. Specifically, CMA finds positive and negative pairs based on the relationship between users and items and conducts contrastive learning on the multimodal features of the constructed pairs. This enables representations of each modality to encode collaborative signals, and features that share common users are placed closer in the embedding space. The explained process is illustrated in Fig. 4.

Contrastive Loss We follow the general concept of contrastive loss from SimCLR [4] and InfoNCE [23] but modify it in a way that reflects the relational structure of the user-item interaction. Given a training sample $(u,i) \in \mathcal{O}$, we consider the multimodal features of item i as an anchor for each modality. Then, we sample one item from $\mathcal{N}_u$, a set of items that user u interacted with, and view its multimodal features as positive instances. Next, we randomly sample multiple negatives from $\mathcal{N}_u^c$, which is the complement set of $\mathcal{N}_u$, and form a set $\mathcal{K}$. Formally, the contrastive loss of visual modality for a training pair (u,i) is as follows:

$$cl_{(u,i,j,\mathcal{K})}^{(v)} = -\log \frac{\exp(\langle e_{i,v}, e_{j,v} \rangle / \tau_v)}{\exp(\langle e_{i,v}, e_{j,v} \rangle / \tau_v) + \sum_{k \in \mathcal{K}} \exp(\langle e_{i,v}, e_{k,v} \rangle / \tau_v)}, \tag{7}$$

where j is a randomly sampled positive item from $\mathcal{N}_u$, and k is one of the randomly sampled negative items from $\mathcal{N}_u^c$. $\langle \cdot, \cdot \rangle$ represents the cosine similarity between the two elements and τ denotes a temperature parameter for each modality. Similarly, we define the contrastive loss for the textual modality as:

$$cl_{(u,i,j,\mathcal{K})}^{(t)} = -\log \frac{\exp(\langle e_{i,t}, e_{j,t} \rangle / \tau_t)}{\exp(\langle e_{i,t}, e_{j,t} \rangle / \tau_t) + \sum_{k \in \mathcal{K}} \exp(\langle e_{i,t}, e_{k,t} \rangle / \tau_t)}. \tag{8}$$

We combine the visual and textual contrastive losses and train the model to minimize the following equation.

$$\mathcal{L}_{CMA} = \sum_{(u,i) \in \mathcal{O}} (cl_{(u,i,j,\mathcal{K})}^{(v)} + cl_{(u,i,j,\mathcal{K})}^{(t)}) \tag{9}$$

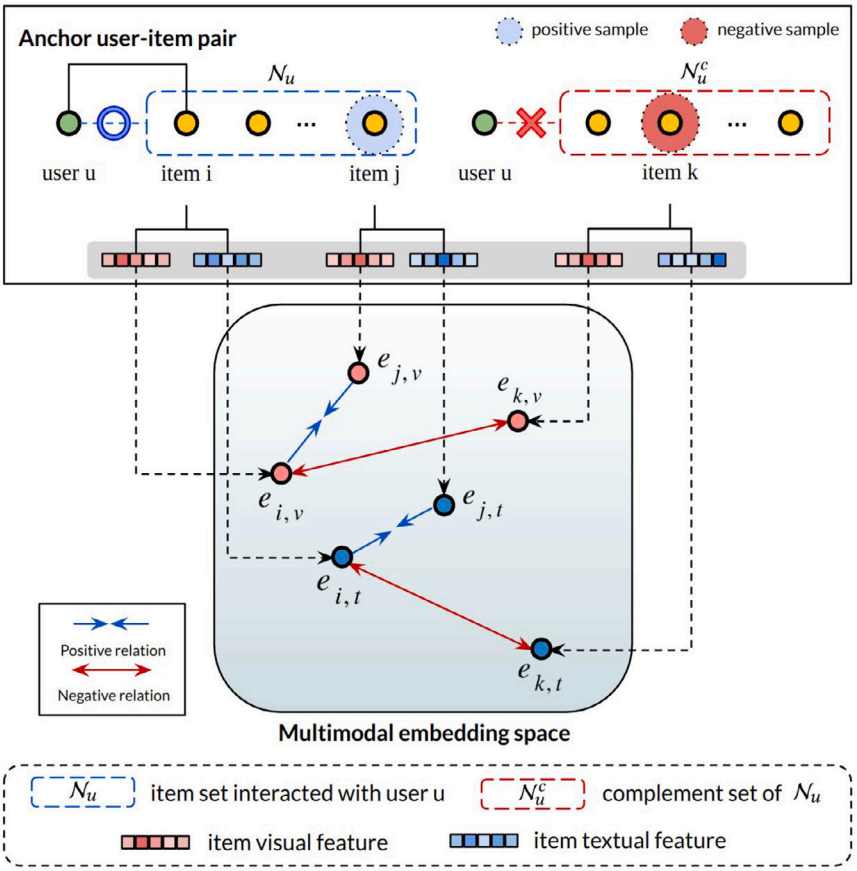


Fig. 4. An illustration of Collaborative Multimodal Alignment (CMA). The positive and negative instances are sampled based on the user-item neighborhood in IMGraph. Item j is sampled as positive because it shares a common user with item i, whereas item k does not. CMA conducts contrastive learning on the multimodal features of the sampled items and encodes collaborative signals. Best viewed in color.

4.3. Multimodal Consistency Regularization

Since the features of each modality stem from different pre-trained encoders, this diversity of the multimodal features may often result in unnecessary variance problems. Therefore, to alleviate the sensitivity and the dependency on certain features or modalities, we add a self-supervised regularization term for the model to be consistent with the views generated from different modalities using graph augmentations.

Graph Augmentation Most existing graph augmentation methods mainly consider homogeneous graphs [26,9]. Here, we present **Mod**ality **Drop** (ModDrop) as a graph augmentation method that applies to our proposed IMGraph as well as other heterogeneous graphs. The idea is to drop out the entire node of a certain modality for each augmentation, reducing dependency on a particular modality. Here, we generate three different augmentations using this paradigm: (1) visual nodes dropout, (2) textual nodes dropout, and (3) multimodal nodes dropout. Augmentation (3) equals $\mathcal{G}_B$, which is a user-item bipartite graph. Regularizing the output consistency with our proposed graph augmentation has several advantages. First, dropping out a modality for each augmentation enables the model to learn more resilient representations that are less dependent on certain modality. Furthermore, our modality dropout enables generalization and over-fitting prevention because the model is trained to be less dependent on specific features or modality. Each augmented graph follows the graph convolution architecture and the prediction layer from Section 4.1.

Consistency Loss The original IMGraph and the augmented graphs generate different representations of the user and item, thus resulting in different predictions. For a training pair (u,i), we calculate the average of the different prediction scores, i.e., $\tilde{y}_{ui} = \frac{1}{S+1}(\hat{y}_{ui} + \sum_{s=1}^{S} \hat{y}_{ui(s)})$, where S is the total number of augmentations. Then, for all training pairs $(u,i) \in \mathcal{O}$, we minimize the L2 distance between the average and each prediction value to increase the consistency:

$$\mathcal{L}_{MCR} = \sum_{(u,i)\in\mathcal{O}} \frac{1}{S+1}\Big(\| \tilde{y}_{ui} - \hat{y}_{ui}) \|_2^2 + \sum_{s=0}^{S} \| \tilde{y}_{ui} - \hat{y}_{ui(s)}) \|_2^2 \Big). \tag{10}$$

Table 2
Statistics of the datasets.

Dataset	#Users	#Items	#Interactions	Density
Amazon	16,138	24,565	114,364	0.00028
Book-Crossing	5,105	5,964	85,722	0.00282
Movielens-1M	6,040	3,260	457,314	0.02323

SMGCN minimizes three loss functions, which are (1) $\mathcal{L}_{BPR}$, a supervised loss for recommendation task, (2) $\mathcal{L}_{CMA}$, a contrastive loss to reinforce the multimodal features to preserve the neighborhood of the user-item graph, and (3) $\mathcal{L}_{MCR}$, a consistency loss that alleviates the variance and dependency problem of multimodal features. We linearly combine the losses with $\lambda_1, \lambda_2 and \lambda_3$ as coefficients that determine the weights of the three losses. The final objective function of the model to optimize is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{BPR} + \lambda_1 \mathcal{L}_{MCR} + \lambda_2 \mathcal{L}_{CMA} + \lambda_3 \parallel \Theta \parallel_2^2, \tag{11}$$

where $\mathcal{L}_{BPR}$ is a BPR loss from Section 4.1 and Θ are the model parameters.

4.4. Complexity analysis of SMGCN

Let $|V|$ and $|E|$ be the number of nodes and edges in the user-item bipartite graph $\mathcal{G}_B$. We connect two multimodal nodes for all item nodes and form the multimodal interaction graph $\mathcal{G}_M$. Then, $\mathcal{G}_M$ does not require additional adjacency, so the space complexity for $\mathcal{G}_M$ is $\mathcal{O}(2|E|)$, which is identical to LightGCN. Since we make 3 graph augmentations for consistency regularization, the total space complexity is $\mathcal{O}(8|E|)$. Accordingly, the total time complexity for the graph convolution becomes $\mathcal{O}(8(|E|d + |I|)Ls\frac{|E|}{B})$, where $|I|$ is the number of items, d is the embedding dimension, L is the total number of convolution layers, s is the number of epochs, and B is the training batch size. For contrastive learning, we perform one matrix multiplication for positive samples and for negatives, and result in the time complexity of $\mathcal{O}((\mathcal{K}+1)B^2d)$. Consistency regularization computes the L2 distance between the results of four graphs and their average, and the time complexity is $\mathcal{O}(4d)$.

5. Experimental setup

5.1. Dataset description

To evaluate the performance of the proposed method, we conduct experiments on three widely used public benchmark datasets for recommendations. The statistics of the datasets are summarized in Table 2. The description of each dataset is as follows:

- **Amazon**[1] contains reviews and metadata of products from Amazon. Among various item categories, we use "Clothing, Shoes and Jewelry" for our experiment based on the intuition that the visual features of the items in this category affects customers' decision more than any other categories. We use explicit purchase histories of users to construct the user-item graph, and representative item images and titles are used as the multimodal information.
- **Book-Crossing**[2] is a dataset containing user ratings on books with values from 1 to 10. We use interactions with rating values higher than 5 for our experiment. The book cover image and the title are used as multimodal information.
- **Movielens-1M**[3] contains people's preferences for movies in the form of implicit ratings from 0 to 5. We only remove the rating value of 0 due to the data sparsity issue. Then manually crawl the poster images of the movies from the web and use them as the visual modality. The names of the movies are used as the textual modality.

For a fair comparison, we use the features extracted from the same pre-trained encoders (ViT, Sentence-BERT) with SMGCN throughout the baselines that incorporate multimodal features.

5.2. Hyperparameter and metrics

We implemented SMGCN with PyTorch and used Xavier [10] initialization for the initial embeddings. The model is optimized using Adam optimizer [15] with learning rate searched within $\{1e^{-4}, 5e^{-4}, 1e^{-3}, 5e^{-3}\}$. For graph-based baseline methods, we set the number of layers as $L = 3$ since a larger number of layers rather deteriorates the performance due to over-smoothing. The model predicts K items with the highest preference for each user, and we report the overall performance comparison with $K = 20$ with the embedding dimension set as 128. Further analysis reports the performance change for the difference in the hyperparameters.

[1] http://jmcauley.ucsd.edu/data/amazon/.
[2] http://www2.informatik.uni-freiburg.de/~cziegler/BX/.
[3] https://grouplens.org/datasets/movielens/1m/.

Table 3
Overall performance comparison with baselines. The best performance is in bold, and the runner-up is underlined. % Improv. represents the relative improvement over the runner-up, expressed as a percentage.

	Amazon		Book-crossing		Movielens-1M	
	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20
MF-BPR	0.1126	0.0881	0.1061	0.0593	0.2656	0.3020
NGCF	0.2019	0.1392	0.0811	0.0429	0.2203	0.2496
LightGCN	0.2193	0.1569	0.0908	0.0498	0.2423	0.2706
SGL	0.2298	0.1911	0.1037	0.0521	0.2674	0.2753
VBPR	0.1554	0.1097	0.0776	0.0492	0.2722	0.3094
MMGCN	0.2287	0.1663	0.1029	0.0577	0.2621	0.2975
MEGCF	0.2351	0.1844	0.1138	0.0581	0.2716	0.3117
MMGCL	0.2402	0.2105	0.1103	0.0602	0.2894	0.2909
SLMRec	0.2421	0.2172	0.1101	0.0537	0.2919	0.3128
MMSSL	0.2514	0.2211	0.1270	0.0632	0.3198	0.3417
SMGCN	**0.2672**	**0.2328**	**0.1311**	**0.0645**	**0.3317**	**0.3597**
% Improv.	+6.28%	+5.29%	+3.23%	+2.06%	+3.72%	+5.27%

5.3. Baselines

To demonstrate the extensive effectiveness of SMGCN, we compare the method with the following state-of-the-art baselines. The baselines are selected in three scopes of research: collaborative filtering, multimodal, and self-supervised learning.

- **MF-BPR** [25] is a representative matrix factorization model optimized with a Bayesian personalized ranking (BPR) loss function. The method assigns latent vectors to each user and item to model their historical interaction.
- **NGCF** [34] is a classical graph-based CF method that applies the message passing algorithm of Graph Convolution Network [38] to the user-item bipartite graph in recommender system framework.
- **LightGCN** [13] simplifies graph convolution in NGCF by removing redundant operations such as non-linear activation and feature transformation that may rather deteriorate the performance in the recommendation task.
- **SGL** [39] introduces self-supervised learning in graph-based CF by contrasting the views of several augmentations (edge dropout, node dropout, and random walk) on the user-item bipartite graph.
- **VBPR** [11] incorporates visual features into the matrix factorization paradigm and learns user-item representations based on their historical interaction. In our experiments, we utilize multimodal features by concatenating them as a single feature vector and inner product with a user ID vector to model their historical interaction.
- **MMGCN** [37] is a graph-based model that captures user preferences on different modalities with a modal-specific user-item bipartite graph. The modal-specific representations are later aggregated with a combination layer. Our experiment uses the element-wise combination method reported in the paper as the best performance.
- **MEGCF** [19] is a recent method that aims to solve the mismatch problem between the multimodal feature extraction and the user interest modeling in the multimodal recommendation. It introduces a collaborative multimodal interaction graph where item nodes are used as a bridge to fuse the collaborative semantics and the multimodal semantics.
- **MMGCL** [40] conducts contrastive self-supervised learning on modal-specific user-item graphs by generating multiple views with modality edge dropout and modality masking.
- **SLMRec** [30] is a recent multimodal recommender system that learns user preferences on different modalities with modal-specific user-item bipartite graphs and conducts augmentation-based contrastive learning on the multimodal features to capture multimodal patterns in the data. We use the model SLMRec-FAC reported in the paper as the best performance.
- **MMSSL** [36] utilizes interactive structure learning and cross-modal contrastive learning to understand the relationship between user-item collaboration and item multi-modal semantics, and to capture the combined effects of user interaction patterns.

6. Experimental results

By conducting various experiments on three benchmark datasets, we aim to answer the following research questions:

- **RQ1**: How does SMGCN perform compared to other state-of-the-art baselines in top-K recommendation task?
- **RQ2**: How does each component of SMGCN contributes to the overall performance?
- **RQ3**: What are the benefits of each modality on the datasets?
- **RQ4**: How sensitive is SMGCN to the changes in the hyperparameters?

6.1. Performance comparison (RQ1)

The experimental results of top-K recommendations are presented in Table 3. The observations are as follows:

Table 4
Component analysis of SMGCN. The best performance is in bold.

	Amazon		Book-crossing		MovieLens-1M	
	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20
Base (LightGCN)	0.2193	0.1569	0.0908	0.0498	0.2423	0.2706
Base + v_{sum}	0.2209	0.1642	0.0895	0.0499	0.2546	0.2745
Base + t_{sum}	0.2221	0.1627	0.0911	0.0528	0.2541	0.2756
Base + vt_{sum}	0.2230	0.1705	0.0913	0.0544	0.2550	0.2780
Base + v_{cat}	0.2241	0.1624	0.0917	0.0538	0.2612	0.2851
Base + t_{cat}	0.2234	0.1611	0.0924	0.0531	0.2547	0.2844
Base + vt_{cat}	0.2281	0.1651	0.0927	0.0554	0.2604	0.2905
Base + v_{node}	0.2257	0.1789	0.1008	0.0521	0.2615	0.2948
Base + t_{node}	0.2211	0.1706	0.1029	0.0564	0.2543	0.2881
IMGraph w/o $g(\cdot)$	0.2314	0.1789	0.1099	0.0580	0.2848	0.3109
IMGraph	0.2399	0.1907	0.1156	0.0592	0.2904	0.3212
IMGraph + $\mathcal{L}_{CMA}$	0.2516	0.2121	0.1205	0.0601	0.3220	0.3512
IMGraph + $\mathcal{L}_{MCR}$	0.2488	0.2050	0.1223	0.0614	0.3192	0.3406
SMGCN	**0.2672**	**0.2328**	**0.1311**	**0.0645**	**0.3317**	**0.3597**

First, we discover that SMGCN's results outperform compared baselines in all datasets and evaluation metrics. Especially, SMGCN improves the previous best result up to 4.41% and 4.20% on average in Recall@20 and NDCG@20, respectively. This advantage is due to the powerful joint representation learning of the features on our IMGraph and the novel Collaborative Multimodal Alignment (CMA) and Multimodal Consistency Regularization (MCR).

Next, the Graph Neural Network (GNN) based methods mostly show promising results in the sparse datasets (Amazon and Book-Crossing). In contrast, conventional factorization methods such as MF-BPR and VBPR are preferable in dense data (Movielens-1M). However, our SMGCN consistently shows the highest performance regardless of the density of the data, indicating the model's robustness on data sparsity. This is mainly because of the sophisticated representation learning on IMGraph, where multimodal node features complement the sparse structure of the user-item bipartite graph. In addition to this, our CMA encodes rich collaborative signals into the multimodal features and supports rich message passing on IMGraph.

Finally, we observe that works that leverage self-supervised learning techniques to enhance the representation learning of the multimodal features shows advanced performances (MMGCL, SLMRec, MMSSL, and SMGCN). However, unlike other methods, the main motivation for using self-supervised learning in SMGCN is to bridge the gap between the user-item CF representations and the multimodal representations for an intricate data fusion. Furthermore, our model includes a novel graph multimodal regularization method that increases robustness in incorporating multimodal features in graph-based CF framework. Combining these self-supervisions, SMGCN achieves precise fusion of collaborative signals and the multimodal features in a graph that acts as a key factor in the state-of-the-art performance.

6.2. Component analysis (RQ2)

In this section, we analyze how each component of SMGCN (Integrated Multimodal Graph (IMGraph), Collaborative Multimodal Alignment (CMA), and Multimodal Consistency Regularization (MCR)) contributes to the overall performance. LightGCN is set as a baseline since it is the simplest form of graph-based CF model, enabling diverse applications to incorporate multimodal features. We first experiment various multimodal feature combination methods including our proposed IMGraph structure. Next, we explore the effects of our two self-supervised optimizations CMA and MCR. Finally, we analyze and compare the results with the performance of our final model SMGCN. The results are summarized in Table 4. Here, we report several crucial observations.

6.2.1. Integrated Multimodal Graph

First, we compare the efficiency of our IMGraph structure with two popular feature combination methods which are summation and concatenation. In Table 4, 'Base + v_{sum}' and 'Base + t_{sum}' is the element-wise summation of visual and textual features with randomly initialized ID embeddings of LightGCN, respectively. 'Base + vt_{sum}' sums up all multimodal features with ID embeddings. Similarly, 'Base + v_{cat}' and 'Base + t_{cat}' concatentates each modality to ID embeddings and 'Base + vt_{cat}' concatenates the whole features. Finally, 'Base + v_{node}' and 'Base + t_{node}' treat each modality feature as a node and connect it to the item ID node in the user-item bipartite graph. Our IMGraph architecture can be interpreted as 'Base + vt_{node}' with edges connecting the visual and textual nodes of the same item. We can observe from the results that integrating the multimodal features generally enhances the performance across all combination methods since additional features act as auxiliary information to distinguish user preference. For all modalities, feature summation showed the smallest increase in the performance among compared methods. When fusing the data of different modalities, feature summation may enhance the overall representation by leveraging the strengths of each source. However, if one source contains noisy information, its contribution could be amplified by summation, resulting in little or no performance gain.

Additionally, we observe the benefits of non-linear feature combination of the multimodal representations by comparing '*IMGraph w/o* $g(\cdot)$' and '*IMGraph*' in Table 4. LightGCN [13] removed non-linearity in the message passing between the user and item ID nodes since the randomly initialized embeddings possess no concrete semantics. However, when combining the multimodal representations in our IMGraph structure, the non-linearity allows the fused representation to capture and express more intricate patterns

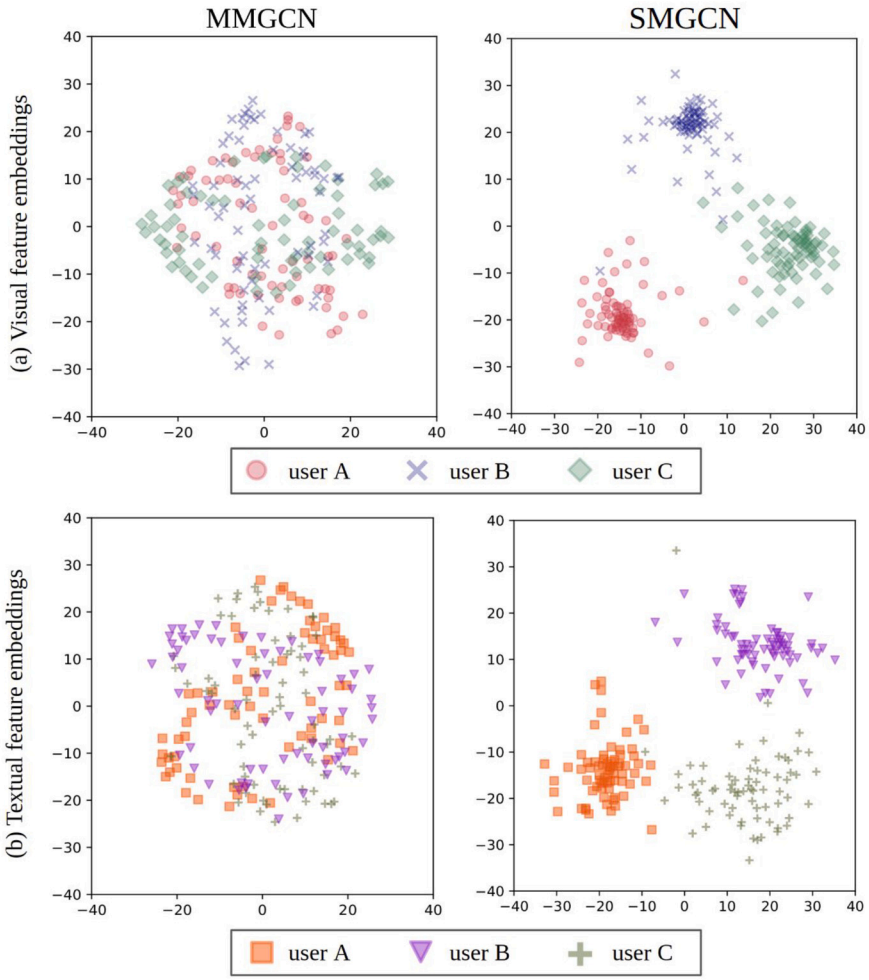


Fig. 5. t-SNE visualization of the multimodal embedding space of Amazon dataset. Each dot represents the multimodal feature of an item. Dots with the same color imply that the items share a common user. We discover that, unlike MMGCN, the features learn the relational structure of the graph and form distinguishable clusters based on the commonly interacted users. Best viewed in color.

and dependencies. As clearly demonstrated in Table 4, the non-linear fusion enhances the recommendation performance across all datasets.

6.2.2. *Collaborative Multimodal Alignment (CMA)*

Through the result of IMGraph + $\mathcal{L}_{CMA}$ in Table 4, we notice that enforcing the multimodal features to gather based on their common interacted users had a significant influence in the overall performance. To see the effect, the t-SNE visualization [31] of multimodal embeddings (i.e., $e_{i,v}^{(0)}$ and $e_{i,t}^{(0)}$ for our SMGCN and learned modal-specific representations for MMGCN) is illustrated in Fig. 5. As we can see in the figure, unlike the multimodal features in MMGCN, features in SMGCN are transformed to form clusters based on the neighborhood of the user-item graph. Our contrastive learning method guides multimodal feature representations to further reflect collaborative signals, thereby propagating much more relevant and fruitful information in the message-passing layers. Additionally, we observe that the features of different modalities of the same items tend to gather near each other, which implies that they are trained to share similar collaborative semantics.

6.2.3. *Multimodal Consistency Regularization (MCR)*

Lastly, our multimodal consistency regularization term $\mathcal{L}_{MCR}$ has beneficial effects on the result when applied solely or with $\mathcal{L}_{CMA}$. This implies that our proposed graph augmentation method ModDrop and optimizing its consistency clearly makes the model robust to various feature incorporation. Specifically, MCR reduces the model dependency on certain modalities and features, eventually contributing to the superior performance of SMGCN.

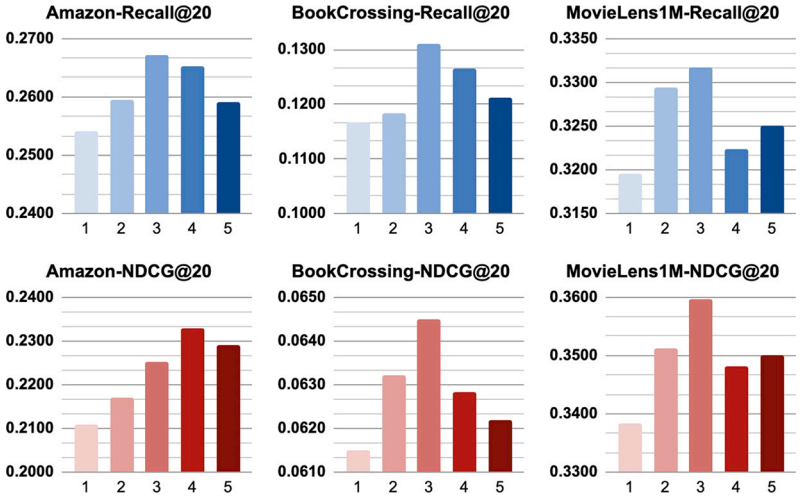


Fig. 6. Impact of the number of graph convolution layers. The X-axis indicates the number of graph convolution layers, and the Y-axis indicates the metrics Recall@20 (Upper row) and NDCG@20 (Lower row).

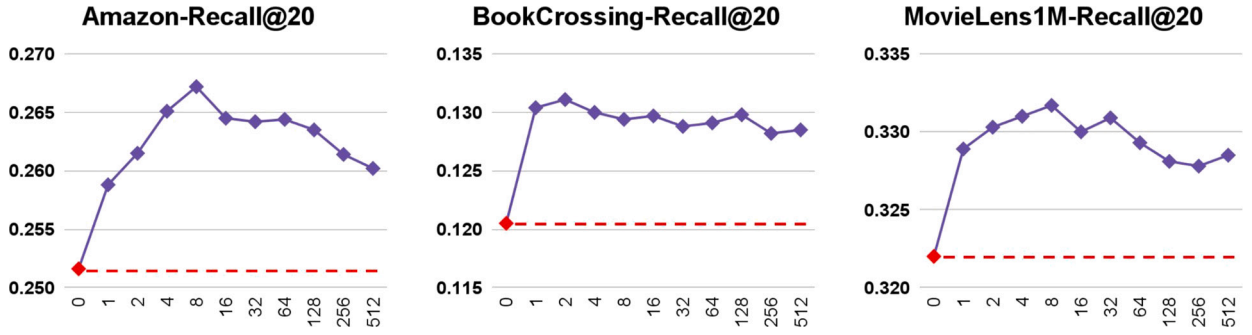


Fig. 7. Impact of the number of negative samples for our multimodal contrastive learning. The red line extending from the leftmost point with zero negative samples indicates the result without contrastive learning. The X-axis indicates the number of negative samples, and the Y-axis indicates the metric Recall@20. Best viewed in color.

6.3. Benfits of each modality (RQ3)

We further analyze the benefits of each modality. The results of 'Base + v_{node}' and 'Base + t_{node}' in Table 4 show how each modality could improve the performance when the modality feature is attached to the user-item bipartite graph as a node in our base model LightGCN. When combined as separate nodes with user-item bipartite graphs, we observe that the multimodal features enhance the recommendation performance since they act as auxiliary information. In Amazon and MovieLens-1M datasets, visual modality had a larger impact. Meanwhile, in Book-Crossing, textual modality had more influence on the result because of the differences in the images across the domains. While the images of items in online shopping platforms (Amazon) and movie posters (MovieLens) are useful side information, the book covers may not be effective for further understanding the contents of books.

6.4. Hyperparameter sensitivity (RQ4)

In this section, we study the effects of hyperparameters on the overall performance of SMGCN. In particular, we aim to find the influences of the following three hyperparameters: 1) the number of graph convolution layers, 2) the number of negative samples for CMA, 3) the temperature coefficients of CMA and 4) the weight values for each loss function.

6.4.1. Model depth analysis

To determine the best value of the number of graph convolution layers, we conduct the experiment on five different layers ranging from one to five. The results are visualized in Fig. 6. The figure demonstrates that three layers of graph convolution generally yield the best result for the model except NDCG@20 in the Amazon dataset. After two or three layers of graph convolution, the node embeddings benefit from the structure of IMGraph by encoding the heterogeneous features of the distant neighbors. However, we can see that exceeding three layers of operations normally degrades performance due to the long-standing over-smoothing problem of graph neural networks.

Fig. 8. A visualization of the performance variations for the combinations of different temperature values. Best viewed in color.

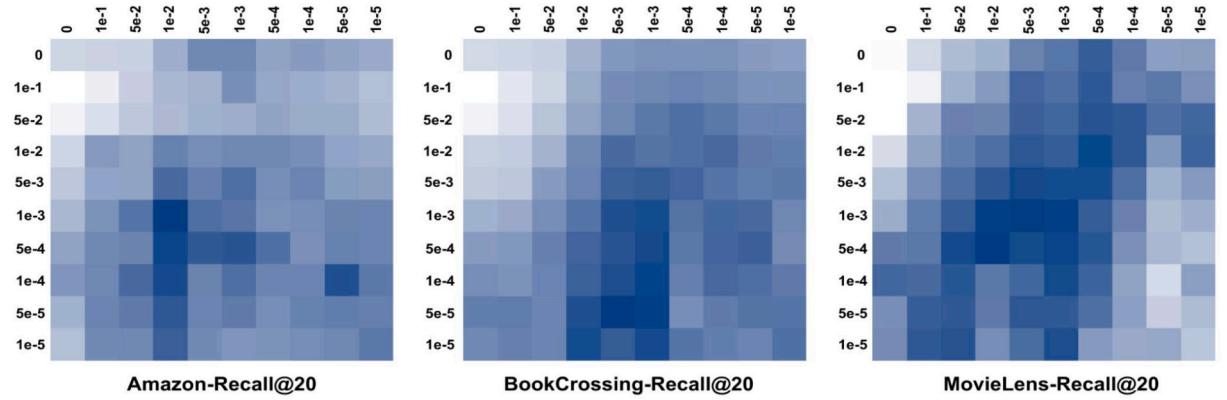


Fig. 9. A visualization of the performance variations for the combinations of the loss weights. The X-axis indicates the weight values for the multimodal contrastive learning, and the Y-axis indicates the values for the consistency regularization. The upper-left grid is the result of *IMGraph* in Table 4, and the first row and column for each image is the result of *IMGraph* $+\mathcal{L}_{CMA}$ and *IMGraph* $+\mathcal{L}_{MCR}$, respectively. Best viewed in color.

6.4.2. Negative samples in CMA

CMA samples positive and negative instances based on the relational structure of user-item interaction history. We conduct experiments with various negative samples within the values {1, 2, 4, 8, 16, 32, 64 128, 256, 512}. To solely observe the impact of the contrastive learning itself, we do not add a consistency regularization term for this experiment. We compare the result of *IMGraph* and *IMGraph* $+\mathcal{L}_{CMA}$ in Table 4, and the results of Recall@20 are visualized in Fig. 7. For Amazon and MovieLens datasets, eight negative samples showed the best result, and for BookCrossing, only two negative samples were enough to benefit from our contrastive learning method. Unlike the conventional contrastive learning in different domains [4], a greater number of negative samples does not necessarily imply greater performance in our setting. We suppose this is due to the distinctiveness of the recommendation datasets compared to others, where many of the instances are co-related to one another based on the interaction history. Therefore, we can observe that considering many instances as negatives has a rather declining impact.

6.4.3. Temperature coefficient of CMA

The temperature coefficient τ in Eq. (7) and Eq. (8) controls the certainty of the model when distinguishing the multimodal features. Therefore, by adjusting the value, we can make the training process either more robust or distinctly classified in the representation space. For our experiment, we chose the temperature by searching within the set $\{0.1, 0.2, 0.5, 1.0\}$. The Recall@20 results with different combinations of the temperature value of each modality are visualized in Fig. 8. Generally, for all datasets, the combinations of relatively lower temperature coefficients yield better performance. This implies that the sharp contrast of the multimodal features in CMA helps distinguish the user preferences in the feature space.

6.4.4. Weight analysis

Our SMGCN updates the parameters by optimizing the linearly combined loss functions as in Eq. (11). In this section, we analyze the impact of the weight values multiplied by the self-supervised losses. We fix the regularization term λ_3 as $1e^{-3}$ and find the optimal values for λ_1 and λ_2. Specifically, we seek the optimal combination of the two weights, each of which is selected within the values $\{1e^{-1}, 5e^{-2}, 1e^{-2}, 5e^{-3}, 1e^{-3}, 5e^{-4}, 1e^{-4}, 5e^{-5}, 1e^{-5}\}$. The results are visualized in Fig. 9, where darker color indicates better performance. We observe some common characteristics from all datasets. First, when the consistency regularization term is solely used, it degrades the result before it reaches a small enough value to pose a beneficial impact. However, the given weights for multimodal contrastive learning instantly increase the performance. Second, the combinations that yield fruitful results, including the highest value, lie between the values {0.01, 0.005, 0.001} for the weight of the multimodal contrastive loss and {0.001, 0.0005, 0.0001} for the weight of the consistency regularization.

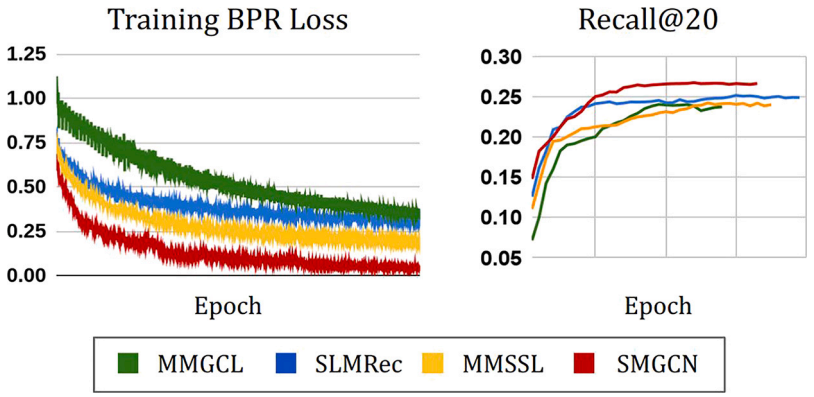


Fig. 10. Comparison of the convergence speed of training BPR loss and Recall@20. Best viewed in color.

6.5. Model training efficiency

In this section we analyze the training efficiency of SMGCN compared to baselines that use self-supervised learning for multimodal recommendation (MMGCL, SLMRec, and MMSSL). Specifically, we compare the convergence speed of training BPR loss and the metric Recall@20. We equally set the learning rate as $1e^{-4}$ for all baselines and use Amazon dataset for this experiment. The results in Fig. 10 clearly demonstrates that SMGCN relatively reaches faster convergence of the training loss and the metric compared to other methods. We speculate this benefit mainly comes from the structure of our proposed IMGraph, where unlike other methods, all the representations are learned and fused in a single graph.

7. Conclusion and future work

We introduce Self-supervised Multimodal Graph Convolutional Network (SMGCN), a novel recommender system that aims to learn cross-modal user preferences over multiple modalities of items with data fusion in a single graph. To achieve this, we first design an efficient graph structure, namely Integrated Multimodal Graph (IMGraph) that enables non-linear integration of the user-item collaborative embeddings and the item multimodal features. Additionally, two novel and efficient self-supervised learning techniques, Collaborative Multimodal Alignment (CMA) and Multimodal Consistency Regularization (MCR), are introduced to enhance multimodal fusion in SMGCN. The experimental results demonstrate the superiority of the proposed model over advanced multimodal models on three benchmark datasets.

For future work, we plan to explore how multiple hops of nodes in IMGraph can affect the performance. The representation learning on IMGraph can be enhanced through higher-order connections of the nodes. Furthermore, we plan to discover how the learned features of SMGCN can be used in various e-commerce tasks such as multimodal search, demand prediction, and advertising.

CRediT authorship contribution statement

Sungjune Kim: Conceptualization, Methodology, Software, Visualization, Writing – original draft. **Seongjun Yun:** Conceptualization, Methodology, Validation. **Jongwuk Lee:** Formal analysis, Writing – review & editing. **Gyusam Chang:** Software, Visualization. **Wonseok Roh:** Software, Visualization. **Dae-Neung Sohn:** Project administration. **Jung-Tae Lee:** Investigation. **Hogun Park:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Sangpil Kim:** Conceptualization, Methodology, Investigation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sangpil Kim reports financial support was provided by NAVER corporation.

Data availability

Data will be made available on request.

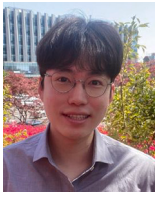
Acknowledgement

This work was supported by NAVER Corp. and supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program (Korea University)).

References

[1] R.v.d. Berg, T.N. Kipf, M. Welling, Graph convolutional matrix completion, arXiv preprint arXiv:1706.02263, 2017.
[2] X. Cai, C. Huang, L. Xia, X. Ren, Lightgcl: simple yet effective graph contrastive learning for recommendation, arXiv preprint arXiv:2302.08191, 2023.
[3] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, T.-S. Chua, Attentive collaborative filtering: multimedia recommendation with item- and component-level attention, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2017, pp. 335–344.
[4] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[5] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: towards visually explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 765–774.
[6] Z. Cheng, S. Han, F. Liu, L. Zhu, Z. Gao, Y. Peng, Multi-behavior recommendation with cascading graph convolution networks, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1181–1189.
[7] H.-g. Chi, M.H. Ha, S. Chi, S.W. Lee, Q. Huang, K. Ramani, Infogcn: representation learning for human skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20186–20196.
[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, in: International Conference on Learning Representations, 2020.
[9] W. Feng, J. Zhang, Y. Dong, Y. Han, H. Luan, Q. Xu, Q. Yang, E. Kharlamov, J. Tang, Graph random neural networks for semi-supervised learning on graphs, Adv. Neural Inf. Process. Syst. 33 (2020) 22092–22103.
[10] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
[11] R. He, J. McAuley, Vbpr: visual Bayesian personalized ranking from implicit feedback, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.
[12] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: Proceedings of the 26th International Conference on World Wide Web, 2017, pp. 173–182.
[13] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, M. Wang, Lightgcn: simplifying and powering graph convolution network for recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 639–648.
[14] Y. Jiang, H. Lin, Y. Li, Y. Rong, H. Cheng, X. Huang, Exploiting node-feature bipartite graph in graph convolutional networks, Inf. Sci. 628 (2023) 409–423.
[15] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[16] Y. Koren, S. Rendle, R. Bell, Advances in collaborative filtering, in: Recommender Systems Handbook, 2021, pp. 91–142.
[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[18] L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10313–10322.
[19] K. Liu, F. Xue, D. Guo, L. Wu, S. Li, R. Hong, Megcf: multimodal entity graph collaborative filtering for personalized recommendation, ACM Trans. Inf. Syst. 41 (2) (2023) 1–27.
[20] Y. Liu, S. Yang, C. Lei, G. Wang, H. Tang, J. Zhang, A. Sun, C. Miao, Pre-training graph transformer with multimodal side information for recommendation, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2853–2861.
[21] L. Lü, M. Medo, C.H. Yeung, Y.-C. Zhang, Z.-K. Zhang, T. Zhou, Recommender systems, Phys. Rep. 519 (1) (2012) 1–49.
[22] A.L. Maas, A.Y. Hannun, A.Y. Ng, et al., Rectifier Nonlinearities Improve Neural Network Acoustic Models, Proc. Icml, vol. 30, Citeseer, 2013, p. 3.
[23] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748, 2018.
[24] N. Reimers, I. Gurevych Sentence-bert, Sentence embeddings using Siamese bert-networks, arXiv preprint arXiv:1908.10084, 2019.
[25] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, 2009, pp. 452–461.
[26] Y. Rong, W. Huang, T. Xu, J. Huang, Dropedge: towards deep graph convolutional networks on node classification, arXiv preprint arXiv:1907.10903, 2019.
[27] X. Su, T.M. Khoshgoftaar, A survey of collaborative filtering techniques, Adv. Artif. Intell. 2009 (2009).
[28] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, K. Zheng, Multi-modal knowledge graphs for recommender systems, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 1405–1414.
[29] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, T.-S. Chua, Mgat: multimodal graph attention network for recommendation, Inf. Process. Manag. 57 (5) (2020) 102277.
[30] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, T.-S. Chua, Self-supervised learning for multimedia recommendation, IEEE Trans. Multimed. (2022).
[31] L. van der Maaten, G. Hinton, Visualizing high-dimensional data using t-629 sne, J. Mach. Learn. Res. 9 (2579–2605) (2008) 630.
[32] P. Veličković, Everything is connected: graph neural networks, Curr. Opin. Struct. Biol. 79 (2023) 102538.
[33] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, Stat 1050 (20) (2017).
[34] X. Wang, X. He, M. Wang, F. Feng, T.-S. Chua, Neural graph collaborative filtering, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 165–174.
[35] X. Wang, H. Jin, A. Zhang, X. He, T. Xu, T.-S. Chua, Disentangled graph collaborative filtering, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1001–1010.
[36] W. Wei, C. Huang, L. Xia, C. Zhang, Multi-modal self-supervised learning for recommendation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 790–800.
[37] Y. Wei, X. Wang, L. Nie, X. He, R. Hong, T.-S. Chua, Mmgcn: multi-modal graph convolution network for personalized recommendation of micro-video, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 1437–1445.
[38] M. Welling, T.N. Kipf, Semi-supervised classification with graph convolutional networks, in: J. International Conference on Learning Representations (ICLR 2017), 2016.
[39] J. Wu, X. Wang, F. Feng, X. He, L. Chen, J. Lian, X. Xie, Self-supervised graph learning for recommendation, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 726–735.
[40] Z. Yi, X. Wang, I. Ounis, C. Macdonald, Multi-modal graph contrastive learning for micro-video recommendation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1807–1811.
[41] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, Z. Huang, Self-supervised learning for recommender systems: a survey. arXiv preprint, arXiv:2203.15876, 2022.
[42] S. Yun, M. Jeong, R. Kim, J. Kang, H.J. Kim, Graph transformer networks, Adv. Neural Inf. Process. Syst. 32 (2019).
[43] F. Zhang, N.J. Yuan, D. Lian, X. Xie, W.-Y. Ma, Collaborative knowledge base embedding for recommender systems, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 353–362.
[44] J. Zhang, Y. Zhu, Q. Liu, M. Zhang, S. Wu, L. Wang, Latent structure mining with contrastive modality fusion for multimedia recommendation, IEEE Trans. Knowl. Data Eng. (2022).
[45] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system: a survey and new perspectives, ACM Comput. Surv. 52 (1) (2019) 1–38.

Sungjune Kim is an integrated Masters/Ph.D. student in the department of artificial intelligence at Korea University. He received his bachelor's degree in business administration at Dongguk University in 2019. His research interests include graph neural networks, relational reasoning, and computer vision.



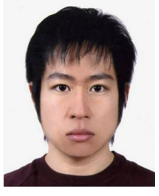
Seongjun Yun is currently working at Amazon in Vancouver. He earned his B.S. in computer science in 2018 from Korea university and Ph.D. in computer science in 2022 from Korea university. His current research interest includes graph neural networks, recommendation systems, and natural language processing.

Jongwuk Lee is currently an associate professor at the Department of Computer Science and Engineering, Sungkyunkwan University (SKKU), South Korea since 2016. He received his B.S. from SKKU in 2006 and Ph.D. from Pohang University of Science and Technology (POSTECH) in 2012. He was an assistant professor at Hankuk University of Foreign Studies (HUFS) in 2014–2016 and a postdoctoral researcher at Pennsylvania State University in 2012–2014. His research interests broadly cover recommender systems, information retrieval, natural language processing, and machine learning algorithm optimization.

Gyusam Chang is a M.S./Ph.D. student at Korea university. He earned his B.S. in electronics and information engineering in 2021 from Korea university, sejong. His research interests includes self-supervised learning, domain generalization, and cross-modal representation learning for computer vision tasks in the field of autonomous driving.

Wonseok Roh is an integrated Masters/Ph.D. student in the department of artificial intelligence at Korea University. He earned his B.S. in computer science in 2022 from the University of Seoul. His research interests include multi-modal detection, autonomous driving, and computer vision.

Dae-Neung Sohn is a computer science expert specializing in shopping search modeling. He earned a B.S. in computer science and a master's degree in natural language processing. His research is focused on query/product feature engineering and ranking modeling for shopping search, as well as fast and accurate NLP models and engines for query reformulation and suggestion. He has published two research papers on content-based mobile spam classification, utilizing stylistically motivated features in Pattern Recognition Letters and ACL. Sohn is currently researching and developing shopping search modeling on Naver corp.



Jung-Tae Lee is a technology executive at Naver Corporation who has extensive experience in information retrieval, recommender systems, and natural language processing. He has authored or co-authored papers on topics such as translation models for online Q&A retrieval, stylistic features for mobile spam filtering, retweet graph analysis for Twitter posts, generative adversarial networks for collaborative filtering, and sentiment-guided deep recommender system. Before joining Naver, he earned a Ph.D. in Computer Science and Engineering from Korea University and worked as research internships at Microsoft Research and Microsoft Research Asia.

Hogun Park is an Assistant Professor in the Department of Computer Science Engineering at Sungkyunkwan University (SKKU). His research interests include relational machine learning, knowledge-based reasoning, explainable AI (XAI), and their applications. Prior to joining SKKU, he worked at the Korea Institute of Science and Technology (KIST) from 2008 to 2013, and for IBM Research-Almaden in 2018 and 2019. Park received his Ph.D. degree in Computer Science from Purdue University in 2020. He has published over 30 fully refereed papers in international journals and conferences in the area of machine learning and data mining.

Sangpil Kim is an assistant professor of Artificial Intelligence at Korea University. His research interests are in 3D computer vision, multi-modal fusion, and generative model. He earned his B.S. degree from Korea University, South Korea in 2015, and his Ph.D. in computer engineering from Purdue University.