Representation Learning with Mutual Influence of Modalities for Node Classification in Multi-Modal Heterogeneous Networks

Jiafan Li[1,2], **Jiaqi Zhu**[1,2,4*], **Liang Chang**[3], **Yilin Li**[1,2], **Miaomiao Li**[1,2], **Yang Wang**[1,2], **Yi Yang**[1] and **Hongan Wang**[1,2]

[1]Institute of Software, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]School of Artificial Intelligence, Beijing Normal University, Beijing, China
[4]Binzhou Institute of Technology, Weiqiao-UCAS Science and Technology Park, Shandong, China
{lijiafan23, limiaomiao22, wangyang223}@mails.ucas.ac.cn, zhujq@ios.ac.cn,
{yilin, yangyi2012, hongan}@iscas.ac.cn, changliang@bnu.edu.cn

Abstract

Nowadays, numerous online platforms can be described as multi-modal heterogeneous networks (MMHNs), such as Douban's movie networks and Amazon's product review networks. Accurately categorizing nodes within these networks is crucial for analyzing the corresponding entities, which requires effective representation learning on nodes. However, existing multi-modal fusion methods often adopt either early fusion strategies which may lose the unique characteristics of individual modalities, or late fusion approaches overlooking the cross-modal guidance in GNN-based information propagation. In this paper, we propose a novel model for node classification in MMHNs, named Heterogeneous Graph Neural Network with Inter-Modal Attention (HGNN-IMA). It learns node representations by capturing the mutual influence of multiple modalities during the information propagation process, within the framework of heterogeneous graph transformer. Specifically, a nested inter-modal attention mechanism is integrated into the inter-node attention to achieve adaptive multi-modal fusion, and modality alignment is also taken into account to encourage the propagation among nodes with consistent similarities across all modalities. Moreover, an attention loss is augmented to mitigate the impact of missing modalities. Extensive experiments validate the superiority of the model in the node classification task, providing an innovative view to handle multi-modal data, especially when accompanied with network structures. The full version including Appendix is available at http://arxiv.org/abs/2505.07895.

1 Introduction

In the real world, numerous online platforms can be characterized as heterogeneous networks [Sun and Han, 2012], which encompass multiple types of nodes connected with various relations, such as movie networks in Douban and product review networks in Amazon [Ni *et al.*, 2019]. With the rapid evolution of Internet, besides textual contents, nodes of certain types also incorporate attributes from other modalities (e.g., images). That enriches the node information and forms multi-modal heterogeneous networks (MMHNs) [Wei *et al.*, 2023; Kim *et al.*, 2023; Chen and Zhang, 2020; Jia *et al.*, 2022], including multi-modal knowledge graphs (MMKGs) [Kannan *et al.*, 2020; Liang *et al.*, 2022; Zhu *et al.*, 2024; Wang *et al.*, 2023b; Sun *et al.*, 2020; Pezeshkpour *et al.*, 2018]. As shown in Figure 1, a movie network comprises movies, actors and directors, with the movie nodes possessing both textual and visual attributes.

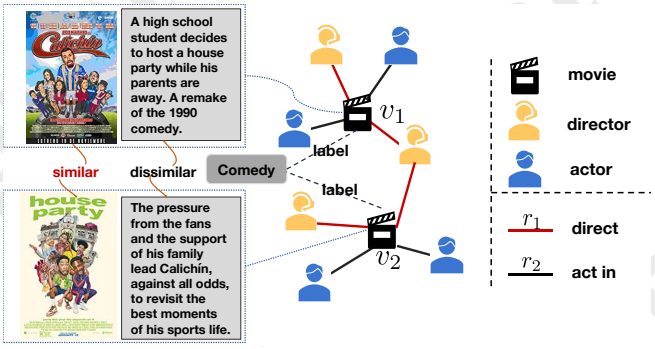

Figure 1: Motivating example of MMHNs and mutual influence of modalities in information propagation for node classification.

It is significant to classify these nodes in the network into distinct categories to understand and analyze key entities, leveraging their multi-modal attributes as well as the network structure [Jangra *et al.*, 2023]. Recently, more and more attentions have been paid to integrate multi-modal features into GNN-based representation learning methods. Some existing studies adopt an early fusion strategy in the encoding process aiming at initial features [Zhang *et al.*, 2019], which may lose the characteristics of individual modalities, while late fusion models learn node embeddings separately for each modality and blend them only at the last layer [Jia *et al.*, 2022].

*Corresponding author

For example in Figure 1, the two film nodes v_1 and v_2 belong to the same category "Comedy", but their textual descriptions differ greatly. Hence, when learning textual embeddings aggregated from neighbors, the similarity-based attentive (2-hop) propagation between them would be limited, potentially resulting in the assignments of them to different categories. Nevertheless, noticing that the images of them are very similar, if that can be acknowledged in determining the propagation weights, the two nodes may eventually own similar embeddings on both modalities and obtain the same label. While, maybe in another scenario, textual similarity plays a pivotal role in realizing a category-oriented aggregation.

Consequently, in order to harness the multi-modal information more effectively in representation learning on MMHNs, it is essential to **consider the mutual influence of modalities during the information propagation process and learn it in an adaptive way**. This poses three main challenges: Firstly, as the influence involves at least two modalities and two nodes, we need to **choose the appropriate granularity to define and distinguish the cross-modal influence**. Here, a tradeoff among expressiveness, intuitive interpretability and model complexity is supposed to be achieved. Secondly, in real-world scenarios, certain types of nodes may have **missing attributes for specific modalities** (e.g., actors and directors lack images), so it is required to handle these incomplete data to ensure a smooth feature propagation process for each modality. Thirdly, even for nodes with attributes across all modalities, some of these attributes may fail to accurately reflect node characteristics. This **misalignment among modalities** would import undesired noise during propagation, causing the learned representations to deviate from correct labels.

To address these issues, this paper proposes a novel model named Heterogeneous Graph Neural Network with Inter-Modal Attention (HGNN-IMA). It aims to learn node representations in MMHNs via capturing the mutual influence of multiple modalities during the information propagation process, thereby supporting the node classification task. Specifically, after pre-processing to encode the attributes of each modality, we devise a heterogeneous network propagation module within the framework of heterogeneous graph transformer, to enrich the node features of each modality by aggregating neighbor information. A key innovation is the nested inter-modal attention mechanism integrated into the classical inter-node attention. When propagating neighbor embeddings to a current node for a specific modality, attention scores are computed as the weighted sum of the similarity-based attention in terms of each modality. The weights are also determined in an attentive and thus adaptive manner, which are further constrained by an attention loss to mitigate the effect of missing modalities. Moreover, these attention scores are necessarily modulated according to the similarity consistency among modalities, amplifying the contribution of modality-aligned nodes. Then, through a feature fusion module, enriched features of all modalities are mixed to generate the final node embeddings. Besides the cross-entropy loss on the fused features, uni-modal features are also incorporated to underline the respective effect of each modality.

In summary, this paper makes the following contributions:

- To the best of our knowledge, this is the first work to adaptively learn and leverage the mutual influence of multiple modalities for model fusion in GNN-based representation learning on MMHNs, tailored for the node classification task.
- A novel model, framed within the heterogeneous graph transformer architecture, is proposed to fulfill the core idea. It features a nested inter-modal attention mechanism on the inter-node attention, plus a modulation term based on similarity consistency to encourage modality alignment, and an additional loss function tackling the modality missing issue. This comprehensive approach accommodates the intricate nature of cross-modal interactions in heterogeneous networks and enhances the category discriminability of node representations.
- Extensive experiments on diverse real-world benchmark datasets demonstrate the effectiveness and stability of the model, achieving significant performance improvements over existing approaches for node classification.

2 Related Work

Heterogeneous networks widely exist in the real world. In the past decade, significant efforts have been devoted to learn node representations with various attention mechanism [Zhuo *et al.*, 2023; Wang *et al.*, 2023a; He *et al.*, 2024; Li *et al.*, 2023b]. For heterogeneous networks with multi-modal attributes, such as texts, images and audios, the fusion of them is necessary. Most of existing models fall into two types, early fusion and late fusion [Jangra *et al.*, 2023; Zhao *et al.*, 2024; Wu *et al.*, 2024; Huang *et al.*, 2024].

Early fusion means the fusion is conducted just after the feature extraction from multi-modal attributes [Wang *et al.*, 2020; Chen and Zhang, 2020]. Besides extending the typical heterogeneous network embedding methods mentioned above by combining initial attributes, HetGNN [Zhang *et al.*, 2019] employs a Bi-LSTM model to encode the feature of each modality and fuse them with mean pooling before intra-type and inter-type aggregations. As this kind of fusion appears prior to information propagation, all modalities of a specific node are merged into one feature, at the expense of losing the characteristics of individual modalities, especially neglecting their different roles in the aggregation process.

Conversely, **late fusion** refers to the fusion after the node representations in terms of each modality have been obtained, so the information propagation of each modality is separately executed [Wei *et al.*, 2019; Tao *et al.*, 2020; Cai *et al.*, 2024; Cao *et al.*, 2022]. As representatives, MHGAT [Jia *et al.*, 2022] performs dual-level aggregations within individual modalities, followed by feature fusion using the modality-level attention mechanism, and FHIANE [Yang *et al.*, 2023a] adds an early fusion module, taking advantage of the consistency and complementarity of multi-modal information. Although these models allow each modality to propagate independently, preserving its original properties, they fail to leverage the information from other modalities when computing the similarity-based attention for accurate aggregation weights. As a result, the learned embeddings may be inconsistent with category labels.

In addition to the two manners of multi-modal fusion above, XGEA [Xu *et al.*, 2023] considers the influence between two modalities during the propagation process, but the role of the current modality itself is neglected and the characteristics of each node in the propagation cannot be distinguished. In contrast, our model explores to adaptively learn and combine the mutual influence among modalities for each node, through a novel nested attention mechanism to enable flexible aggregations.

Beyond that, IDKG [Li *et al.*, 2023a] regards the knowledge graph as another modality. Its translation-based embeddings, along with visual and textual features, are fused into a unified representation, so is able to solve the node classification problem with a different usage of network structure.

3 Problem Formulation

Definition 1 (Multi-Modal Heterogeneous Networks)**.** *A Multi-Modal Heterogeneous Network (MMHN) is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{M})$, where $\mathcal{V}$, $\mathcal{E}$ and $\mathcal{M}$ respectively denote the set of nodes, edges and multi-modal attributes (e.g., texts, images and audios). There are also a node type mapping function $\psi(v) : \mathcal{V} \to \mathcal{O}$ and an edge type mapping function $\phi(e) : \mathcal{E} \to \mathcal{R}$, where $\mathcal{O}$ and $\mathcal{R}$ denote the set of pre-defined types of nodes and edges respectively, with $|\mathcal{O}| + |\mathcal{R}| > 2$.*

Notice that although each node in $\mathcal{V}$ can contain multiple types of attribute information from different modalities in $\mathcal{M}$., some modalities are not available for certain node types. For example, reviews in Amazon are probably not associated with an image. Thus, we define a mapping function $f(o) : \mathcal{O} \to 2^{\mathcal{M}}$ to indicate the subset of modalities that each node type possesses. In this context, for each node $v_i \in \mathcal{V}$, its attribute of each modality $m \in f(\psi(v_i))$ is denoted as x_i^m.

It is reasonable to assume there is only one target node type $o^* \in \mathcal{O}$ required to be categorized, such as films in movie networks and items in product review networks. Given a pre-defined set of categories $\mathcal{C}$, the node classification task on MMHNs aims to assign a category label $y_i \in \mathcal{C}$ to each target node v_i in a semi-supervised setting.

4 Proposed Model HGNN-IMA

In this section, we at first present the framework of the proposed model, and then elaborate on the key modules.

4.1 Framework

HGNN-IMA is structured into three modules: pre-processing module, heterogeneous network propagation module, and feature fusion module. Since most of real-world MMHNs have only text and vision attributes, we illustrate our model in Figure 2 with these two modalities, i.e., $\mathcal{M} = \{\mathrm{T}, \mathrm{V}\}$.

As pre-processing, the feature for each modality of each node is extracted through a corresponding pre-trained encoder. Then, the core propagation module is designed to learn enriched representations of nodes in the framework of heterogeneous graph transformer (HGT) [Hu *et al.*, 2020]. To capture synergetic effect of multi-modal attributes, we primarily innovate with a **Cross-modal Influence Unit**, which nests the inter-modal attention into modal-specific inter-node attentions, allowing the aggregation weights to be determined by all modalities together. Besides, an additional loss for inter-modal attention and a modulation on inter-node attention are employed during propagation to mitigate the impact of nodes with missing and misaligned modalities respectively.

At last, the feature fusion module incorporates the representations from different modalities also in an attentive manner. For semi-supervised training, the cross-entropy loss on both multi-modal and uni-modal features, as well as the attention loss are combined with equal proportions.

4.2 Heterogeneous Network Propagation Module

In order to aggregate the embeddings of heterogeneous neighbors for the semantic enrichment of each node, we adopt HGT as the base architecture instead of dual-level ones [Hu *et al.*, 2019; Wang *et al.*, 2019] to uniformly handle various types of nodes and edges with the help of multiple type-dependent feature projection functions. That facilitates the characterization of complicated interactions among heterogeneous entities and their multi-modal attributes.

Considering the discrepancy among node types, for each node v_i and its feature on a modality m, we initially make a linear function dependent on node type to get the embedding $\mathbf{h}_i^{(0),m}$ at the 0-th layer, as the starting point of aggregation.

$$\mathbf{h}_i^{(0),m} = l^{\mathrm{I}}_{\psi(v_i)}(\bar{\mathbf{h}}_i^{(0),m}) \tag{1}$$

where $\bar{\mathbf{h}}_i^{(0),m}$ is the output of the m-encoder in the pre-processing phase. Then, each node v_i expects to receive information from its neighbor set, denoted as $N_i \subseteq \mathcal{V}$.

Cross-modal Influence Unit

Like previous work of multi-modal fusion in GNNs, the propagation here is still executed for each modality m (e.g., text or vision) separately, but as exemplified in Figure 1, other modalities should also take effect in deciding the importance (weights) of neighbors. Thus, we need to compute these weights from the perspective of each influencing modality based on the similarity of the features on that modality, and unequally combine them in an adaptive way.

Specifically, for each influencing modality $m' \in \mathcal{M}$, given a current node v_i and one of its neighbors $v_j \in \mathcal{N}_i$, the importance of v_j for v_i at the k-th layer can be calculated as the similarity of the two nodes on this modality, and realized by a bilinear transformation following HGT [Hu *et al.*, 2020]:

$$g_{ij}^{(k),m'} = l^{\mathrm{K}}_{\psi(v_j)}(\mathbf{h}_j^{(k-1),m'}) \cdot W^{\mathrm{NODE}}_{\phi(v_j,v_i)} \cdot l^{\mathrm{Q}}_{\psi(v_i)}(\mathbf{h}_i^{(k-1),m'}) \tag{2}$$

where l^{K} and l^{Q} are two linear functions dependent on the node type similar to l^{I}, and W^{NODE} is a learnable matrix dependent on the edge type. Then, the inter-node attention score between v_i and its neighbor v_j at the k-th layer on the modality m' is denoted as $\alpha_{ij}^{(k),m'}$ and computed as follows:

$$\alpha_{ij}^{(k),m'} = \underset{\forall j \in N_i}{\mathrm{softmax}}\left(g_{ij}^{(k),m'}\right) = \frac{\exp\left(g_{ij}^{(k),m'}\right)}{\sum_{j' \in N_i} \exp\left(g_{ij'}^{(k),m'}\right)} \tag{3}$$

Here, a single attention head is used for simplicity, and is easy to be extended to multiple heads similar to HGT.

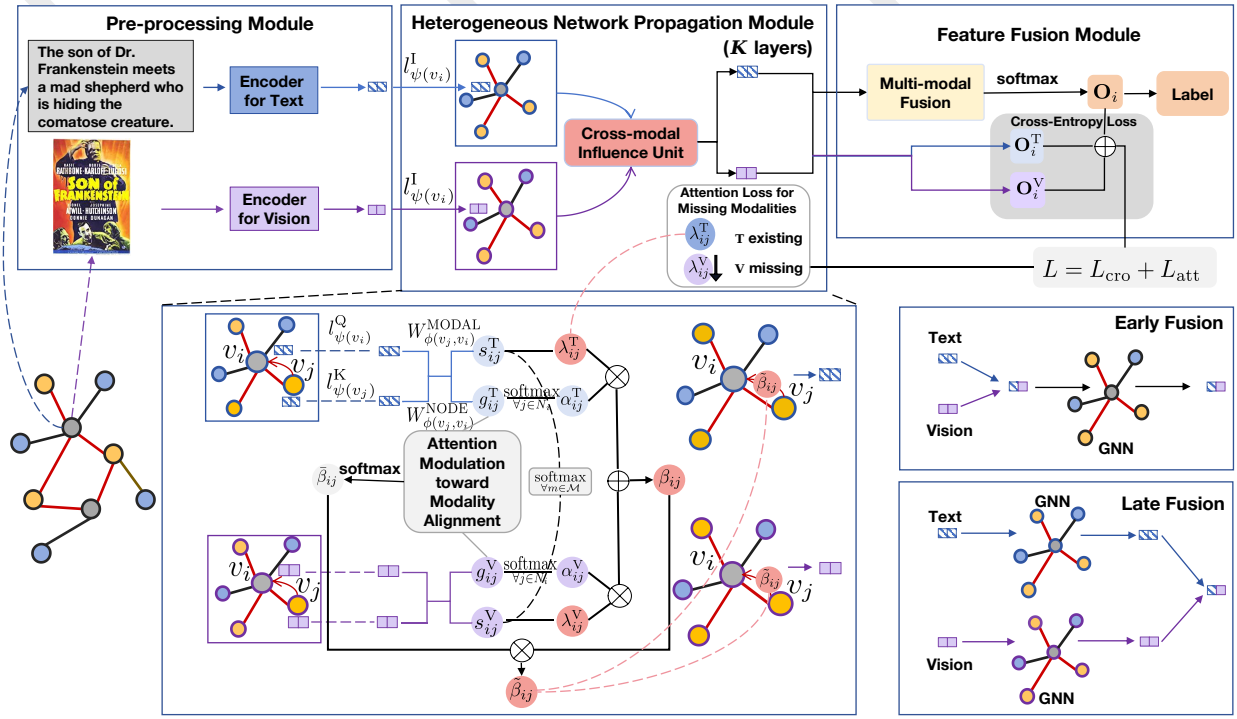

Figure 2: Framework of HGNN-IMA comprising of three modules, compared with existing modal fusion strategies on the bottom right.

With these attention scores on each modality, it is crucial to determine which modalities are suited to play more important roles. That naturally depends on the features of the two nodes v_i and v_j, so should be learned adaptively. To this end, we design a nested attention mechanism. At first, the inter-modal attention score for the modality m' is calculated in two steps:

$$s_{ij}^{(k),m'} = l_{\psi(v_j)}^{\mathrm{K}}(\mathbf{h}_j^{(k-1),m'}) \cdot W_{\phi(v_j,v_i)}^{\mathrm{MODAL}} \cdot l_{\psi(v_i)}^{\mathrm{Q}}(\mathbf{h}_i^{(k-1),m'}) \tag{4}$$

$$\lambda_{ij}^{(k),m'} = \underset{\forall m' \in \mathcal{M}}{\mathrm{softmax}}\left(s_{ij}^{(k),m'}\right) = \frac{\exp\left(s_{ij}^{(k),m'}\right)}{\sum_{m'' \in \mathcal{M}} \exp\left(s_{ij}^{(k),m''}\right)} \tag{5}$$

Notice that we use the same features of the two nodes as the inter-node attention α, but different parameters W^{MODAL} to characterize their correlations from the perspective of cross-modal influence, instead of similarity.

Next, we apply the inter-modal attention $\lambda_{ij}^{(k),m'}$ in Equation 5 on the modal-specific inter-node attention $\alpha_{ij}^{(k),m'}$ in Equation 3, and obtain the combined inter-node attention $\beta_{ij}^{(k)}$ reconciling the features and effects of all modalities:

$$\beta_{ij}^{(k)} = \underset{\forall j \in N_i}{\mathrm{softmax}}\left(\sum_{m'=1}^{\mathcal{M}} \left(\lambda_{ij}^{(k),m'} \alpha_{ij}^{(k),m'}\right)\right) \tag{6}$$

It is worth noting that this attention is independent of the current (influenced) modality m. In other words, for a specific node, the features of all modalities are propagated according to unique weights from its neighbors. Although compromising the expressiveness to some extent, this simplification is reasonable to maintain a moderate number of parameters and highlight the essential idea of the cross-modal influence, which is certified via ablation study in Section 5.5.

In view of the complexity of multi-modal data, there exist the phenomenons of modality misalignment and even missing. Therefore, the inter-modal attention and the inter-node attention need to be adjusted accordingly to fulfill a smooth and category-oriented propagation.

Attention Loss for Missing Modalities

As formulated in Section 3, there exist some types of nodes not possessing all modalities originally, and their features on the missing modalities have to be completed in some way, so they should not take much effect compared to modalities with real attributes. To address this issue, we specially design a loss function to constrain the inter-modal attention scores computed by Equation 5 in such cases are not too large:

$$L_{\text{att}} = \frac{1}{K \cdot |\mathcal{M}|} \sum_{v_i \in \mathcal{V}} \sum_{v_j \in N_i} \sum_{1 \le k \le K} \sum_{m' \notin f(\psi(v_j))} \lambda_{ij}^{(k),m'} \tag{7}$$

where K is the number of propagation layers.

Attention Modulation toward Modality Alignment

Even for those nodes possessing all modalities, the attributes of some modalities may not represent the correct meaning of nodes due to the irregularity of Internet. If a node suffering such misalignment propagates too much information to the current node, it is inevitable to import noises in the aggregation. Since the similarity of two nodes on each modality

has been computed in Equation 2, we can utilize the consistency of these scores to imply the modality alignment degree of each neighbor, and thus get a new inter-node attention $\bar{\beta}$ for the aggregation weights as follows.

$$\bar{\beta}_{ij}^{(k)} = \underset{\forall j \in N_i}{\text{softmax}} \left(\sum_{m_1, m_2 \in \mathcal{M}} |g_{ij}^{(k),m_1} - g_{ij}^{(k),m_2}| \right) \tag{8}$$

Then, through mixing the two inter-node attentions from different views into final aggregation weights $\tilde{\beta}_{ij}^{(k)}$, the embedding of each modality m at the k-th layer is computed as:

$$\tilde{\beta}_{ij}^{(k)} = \underset{\forall j \in N_i}{\text{softmax}}(\beta_{ij}^{(k)} \cdot \bar{\beta}_{ij}^{(k)}) \tag{9}$$

$$\tilde{\mathbf{h}}_i^{(k),m} = \sum_{j \in N_i} \tilde{\beta}_{ij}^{(k)} \cdot l^{\text{M}}_{\psi(v_j)}(\mathbf{h}_j^{(k-1),m}) \cdot W^{\text{MSG}}_{\phi(v_j,v_i)} \tag{10}$$

where l^{M} is the fourth function dependent on node type and W^{MSG} is the third learnable matrix dependent on edge type.

At last, to avoid over-smoothing, we introduce a residual connection to get the final output at the k-th layer, with the sigmoid function and another type-dependent function l^{A}:

$$\mathbf{h}_i^{(k),m} = l^{\text{A}}_{\psi(v_i)} \left(\sigma(\tilde{\mathbf{h}}_i^{(k),m}) \right) + \mathbf{h}_i^{(k-1),m} \tag{11}$$

4.3 Feature Fusion Module

Although the computation of embeddings for each modality has already taken cross-modal influence into accounts, they are still required to be fused to get final representations. We use standard modality-level attention to learn adaptive importance of each modality for classification at the last layer:

$$\omega_i^m = \mathbf{w_2} \cdot \tanh(\mathbf{w_1} \cdot \mathbf{h}_i^{(K),m}) + b_2 \tag{12}$$

$$\delta_i^m = \text{softmax}(\omega_i^m) = \frac{\exp(\omega_i^m)}{\sum_{m'' \in \mathcal{M}} \exp(\omega_i^{m''})} \tag{13}$$

The final embedding $\mathbf{Z}_i$ of node v_i is then calculated, and the probability distribution $\mathbf{O}_i$ of v_i for each category is obtained:

$$\mathbf{Z}_i = \sum_{m \in \mathcal{M}} \delta_i^m \cdot \mathbf{h}_i^{(K),m} \tag{14}$$

$$\mathbf{O}_i = \text{softmax}(\mathbf{W}_1 \cdot \mathbf{Z}_i) \tag{15}$$

4.4 Training Objective

For semi-supervised classification, the cross-entropy loss is used. To embody the effect of each modality, we incorporate the losses of individual modalities into the fused one:

$$L_{\text{cro}} = \frac{1}{1 + |\mathcal{M}|} \left(\sum_{v_i \in \mathcal{V}_{\text{L}}^*} \mathbf{y}_i^\top \cdot \log(\mathbf{O}_i) + \sum_{m \in \mathcal{M}} \sum_{v_i \in \mathcal{V}_{\text{L}}^*} \mathbf{y}_i^\top \cdot \log(\mathbf{O}_i^m) \right) \tag{16}$$

where $\mathcal{V}_{\text{L}}^*$ is the labeled target node set, and $\mathbf{y}_i$ is the one-hot label vector for v_i. $\mathbf{O}_i^m$ is the embedding for each modality m, computed similar to $\mathbf{O}_i$ in Equation 15, but replacing $\mathbf{Z}_i$ by $\mathbf{h}_i^{(K),m}$. Then, the whole loss function is expressed as the equal-weight sum of the two losses:

$$L = L_{\text{cro}} + L_{\text{att}} \tag{17}$$

It can be inferred that the total computational complexity is $O(|\mathcal{V}|^2 \cdot |\mathcal{M}|^2)$, which demonstrates the scalability of the model when handling large graphs and multiple modalities. To save space, the detailed analysis is put in Appendix.

Dataset	Nodes	Edges	Edge types	Categories
DOUBAN	6627	15032	4	2
IMDB	11616	34212	4	3
AMAZON	13189	174154	3	3
AMAZON-1	58088	632238	4	12
AMAZON-2	58088	632238	4	12

Table 1: Dataset statistics.

5 Experiments

In this section, we first introduce the datasets, baselines, and experimental settings. Subsequently, we demonstrate the superiority of our model over baselines on node classification, followed by ablation study and hyper-parameter analysis.

The experiments were performed on NVIDIA Tesla V100 32 GPUs, and implemented in Python 3.9 with PyTorch.[1]

5.1 Datasets

We evaluate our model on five diversified real-world benchmark datasets. The statistics of them are shown in Table 1.

- DOUBAN[2] and IMDB[3] collect data from two online movie websites respectively. Movies (M), actors (A) and directors (D) compose the heterogeneous network with edge types AM, MA, MD and DM. Besides textual descriptions, each movie possesses a poster image, and can be categorized into "Action", "Comedy" and "Drama" ("Drama" only exists in IMDB).
- AMAZON[4] contains items (I) and user reviews (U) in Amazon, with three types of relations: UI, IU and II. Each item (product) has a text and an image. Three subcategories of appliances are used for classification.
- AMAZON-1 and AMAZON-2 are self-constructed larger datasets under the Electronics category in the AMAZON dataset[5]. Four types of relation between items are selected: also_buy, also_viewed, buy_after_viewing, and bought_together. In AMAZON-2, the price of items is added as the third modality.

5.2 Baselines

For baselines, we choose representative models of the three manners regarding modalities: (1) typical heterogeneous graph neural networks without handling multi-modal attributes (**HAN** [Wang *et al.*, 2019], **SHGP** [Yang *et al.*, 2022], **SeHGNN** [Yang *et al.*, 2023b], **HERO** [Mo *et al.*, 2024]) and **HGT** [Hu *et al.*, 2020], for which we employ *both early fusion and late fusion* strategies; (2) special heterogeneous graph neural networks tackling multi-modal attributes (**HetGNN** [Zhang *et al.*, 2019] by *early fusion* , **MHGAT** [Jia *et al.*, 2022] by *late fusion* with two versions of inter-node aggregation, and **XGEA** [Xu *et al.*, 2023] considering *fixed*

[1]The code is available at https://github.com/Jiafan-ucas/HGNN-IMA

[2]https://github.com/jiaxiangen/MHGAT/tree/main/douban

[3]https://github.com/Jhy1993/HAN/tree/master/data/imdb

[4]https://github.com/jiaxiangen/MHGAT/tree/main/amazon

[5]https://nijianmo.github.io/amazon/index.html

Datasets		DOUBAN		IMDB		AMAZON		AMAZON-1		AMAZON-2	
Metrics		Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HAN	early	0.8707	0.8666	0.7267	0.7262	0.8594	0.8015	0.8532	0.6866	0.8542	0.6927
	late	0.8737	0.8699	0.7300	0.7286	0.8337	0.7737	0.8250	0.6157	0.8208	0.5936
SHGP	early	0.8319	0.8288	0.5488	0.5447	0.7483	0.6344	0.5989	0.3311	0.5678	0.3038
	late	0.8224	0.8256	0.5320	0.5180	0.7748	0.6920	0.5911	0.3205	0.5844	0.3255
SeHGNN	early	0.8667	0.8652	*0.7496*	*0.7478*	0.8726	0.8289	0.8554	0.7323	0.8522	0.7561
	late	0.8677	0.8624	0.7453	0.7438	0.8550	0.8122	0.8571	0.7638	0.8554	0.7660
HERO	early	0.8533	0.8493	0.6517	0.6102	0.8295	0.7699	0.8023	0.6755	0.8058	0.6546
	late	0.8283	0.8252	0.6936	0.6848	0.8207	0.7547	0.8136	0.6862	0.8012	0.6723
HGT	early	0.8508	0.8483	0.7407	0.7381	*0.8773*	*0.8302*	0.8883	0.7682	*0.8882*	*0.7807*
	late	0.8654	0.8629	0.7419	0.7407	0.8703	0.8212	*0.8931*	*0.7990*	0.8871	0.7799
HetGNN (early)		0.8366	0.8332	0.5068	0.4906	0.8328	0.7636	0.7012	0.5187	0.7129	0.4977
MHGAT (late)	max	0.8629	0.8574	0.7364	0.7249	0.8638	0.8084	0.8011	0.6729	0.8003	0.6486
	sum	0.8545	0.8468	0.7220	0.7127	0.7963	0.6734	0.7975	0.5929	0.7873	0.5433
IDKG		0.8462	0.8451	0.7410	0.7387	0.8752	0.8276	0.8504	0.5164	0.8604	0.5268
XGEA		*0.8765*	*0.8728*	0.7126	0.7047	0.8596	0.8001	0.8847	0.7226	0.8872	0.7301
HGNN-IMA		**0.8778**	**0.8758**	**0.7578**	**0.7560**	**0.8870**	**0.8427**	**0.8946**	**0.8233**	**0.8905**	**0.8182**

Table 2: Overall results of HGNN-IMA and baselines on five datasets by two metrics. The best scores are in bold and the second in italic.

influence between modalities); (3) translation-based methods treating the network structure a new modality (**IDKG** [Li *et al.*, 2023a]). The details are provided in Appendix.

5.3 Experimental Settings

As to AMAZON, IMDB and DOUBAN datasets, we directly use the encoded features provided by MHGAT [Jia *et al.*, 2022]. For self-constructed datasets AMAZON-1 and AMAZON-2, texts and images are encoded using CLIP [Radford *et al.*, 2021], while price is embedded through an FFNN. For missing visual attributes, we complete them with the text features after encoding, as the only available information.

We divide each dataset into the training set (20%), the validation set (10%), and the test set (70%). The model is trained using the Adam optimizer with a learning rate of 0.001, incorporating 3 layers of propagation ($K = 3$) and setting the maximum number of iterations as 300. Each layer's output is subjected to a dropout with a rate of 0.6, and the dimension d of all embeddings is standardized to 64. During propagation, we employ multi-head attention with the number of heads set to 8. To further prevent over-fitting, we employ an early stopping mechanism with a patience of 50, which activates when the validation loss exceeds any previously recorded values and its accuracy dips below the highest recorded one. Both Micro-F1 and Macro-F1 are used for evaluation. We report the average results of five executions with different seeds.

5.4 Overall Results

Table 2 shows the overall results on five datasets. It can be seen that our model consistently outperforms all the baselines across all datasets. For the first three datasets, there are 1.2%, 1.6%, and 0.3% gains in Macro-F1 compared to the strongest baseline. While, for the two larger datasets with more categories, the Macro-F1 value is promoted by around 2.5%. That indicates the model is able to achieve a more balanced performance improvement across all categories on larger datasets.

Specifically, HGNN-IMA is superior to typical HGNNs such as HAN, SHGP, HERO and SeHGNN to a great extent, no matter early pre-processing or late post-processing for multiple modalities, so the special treatment on multi-modal attributes in MMHNs is necessary to understand the categories of entities and thereby worth studying. The second-best performances of HGT on most datasets are attributed to its use of type-dependent parameters to capture heterogeneous attentions over each edge from a global view. Our model further enhances HGT via effectively handling modalities.

Compared to HetGNN with explicit early concatenation of multi-modal attributes, as well as MHGAT employing late modality-level attention mechanism, the notable promotion of our model can naturally give credit to the innovative nested attention mechanism. It adaptively learns the cross-modal influence and determines the aggregation weights of each node to attain expected information propagation, rather than just blending multi-modal features before or after the propagation. Although XGEA considers the influence of another modality in propagation, it assumes a fixed influence relation and neglects the role of the current modality itself, so fails to identify the complicated synergy of modalities on category-oriented propagation for each node. Also, the advantage over IDKG implies GNN-based propagation is essential to learn discriminative node embeddings through incorporating high-order correlations, especially facing multi-modal attributes.

Besides, we calculate the standard deviation (STD) of our model and sub-optimal HGT, and conduct a two-sample t-test between them, presented in Table 3. The results confirm the stability of our model. Additionally, the p-value is less than 0.05 in all datasets, indicating a significant difference be-

Metrics	DOUBAN	IMDB	AMAZON	AMAZON-1	AMAZON-2
Ours (STD)	0.0015	0.0031	0.0019	0.0021	0.0033
HGT (STD)	0.1261	0.0035	0.0072	0.0106	0.0076
t-value	3.01	4.18	4.09	4.36	6.14
p-value	0.039	0.014	0.015	0.012	0.004

Table 3: Standard deviation and t-test for HGNN-IMA and HGT.

tween HGNN-IMA and baseline models. The intuitive comparison on learned embeddings is displayed in Appendix.

5.5 Ablation Study

Here, we design hierarchical variants of HGNN-IMA to inspect the importance of the key components in the model. Other ablative models are analyzed in Appendix.

- Removing or changing the Cross-modal Influence Unit
  - HGNN-IMA−cross: It eliminates this unit and directly uses inter-node attention $\alpha_{ij}^{(k),m'}$ to replace $\beta_{ij}^{(k)}$ in Equation 6 for the aggregation.
  - HGNN-IMA−adapt: It utilizes the mean of $\alpha_{ij}^{(k),m'}$ on each modality to serve as $\beta_{ij}^{(k)}$ in Equation 6, losing weight adaptability.
  - HGNN-IMA+inf: It distinguishes the influenced modality m when computing the inter-modal attention to form $\lambda_{ij}^{(k),m',m}$ in Equation 5.
  - HGNN-IMA−nei: It neglects the distinction among the neighbors v_j when computing the inter-modal attention to form $\lambda_{i}^{(k),m'}$ in Equation 5.
- Removing attention modulation for modality alignment
  - HGNN-IMA−align: It directly uses inter-node attention $\beta_{ij}^{(k)}$ to replace $\tilde{\beta}_{ij}^{(k)}$ in Equation 10.
- Removing some part of loss functions
  - HGNN-IMA−L_{att}: It removes the attention loss to disregard the modality missing issue.
  - HGNN-IMA−L_{ind}: It removes the individual modality part from the cross-entropy loss.

Table 4 proves that the Cross-modal Influence Unit is beneficial, and the weights of influence should be adaptively learned rather than pre-defined. We also find compared to the traditional attention score $\alpha_{ij}^{(k),m}$, our nested one $\beta_{ij}^{(k)}$ aligns more closely with node categories, which is visualized in Appendix. Regarding the specific design of the nested inter-modal attention, when altering the granularity such as adding the influenced modality or removing the influencing node, the performance declines in varying degrees. Hence, while the core idea is simple, the tradeoff between expressiveness and complexity in defining the cross-modal influence is nuanced.

Then, when removing the modulation term, the model exhibits decreased performance for all datasets. That certifies the unconstrained propagation according to the cross-modal influence is susceptible to noises imported by inaccurate modality attributes. The modulation based on similarity

Variants	DOUBAN	IMDB	AMAZON	AMAZON-1	AMAZON-2
-cross	*0.8661*	0.7344	0.8173	0.8125	0.8067
-adapt	0.8594	0.7360	*0.8397*	0.8044	0.7696
+inf	0.8614	0.7319	0.8332	0.8101	0.7798
-nei	0.8599	0.7256	0.8241	0.8082	0.7921
-align	0.8562	*0.7487*	0.8389	0.8056	0.8095
-L_{att}	0.8467	0.7436	0.8334	\	\
-L_{ind}	0.8602	0.7469	0.8392	*0.8218*	**0.8270**
Ours	**0.8758**	**0.7560**	**0.8427**	**0.8233**	*0.8182*

Table 4: Macro-F1 of ablative models. The best scores are marked in bold and second in italic. "\" means inapplicable due to full data.

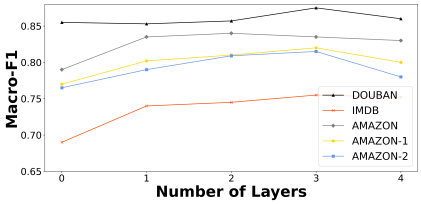

Figure 3: Macro-F1 values with varied layer numbers

consistency just weakens this awkward impact through focusing more on those nodes which realize modality alignment.

At last, the ablation for loss functions highlights their respective roles within the model. An exception appears when the single-modality loss is removed in the AMAZON-2 dataset. That can be explained as the introduction of the new modality, so it should be cautious to consider the contribution of individual modalities, which may bring a negative impact.

5.6 Hyper-parameter Analysis - Layer Number

We change the layer number K from 0 to 4, as shown in Figure 3. It can be observed that the trend of Macro-F1 scores is similar across all datasets and arrives an optimal value when $K = 3$. That is because adopting fewer than 3 layers prevents sufficient information exchange, whereas the continual increase of layers would lead to over-smoothing.

6 Conclusion

This paper delves into the intricate problem of node representation learning within multi-modal heterogeneous networks, characterized with complicated interactions of modalities and node/edge types. To overcome the limitations associated with early or late fusion of multi-modal features, we put the fusion inside the GNN-based propagation process, thereby prompting node representations to align closely with category labels. Notably, the innovative inter-modal attention acting on the modal-specific inter-node attention is proposed to enable adaptive modal fusion, based on the heterogeneous graph transformer framework. Moreover, another two critical factors in multi-modal data, modality alignment and missing, are also integrated into the model in a straightforward way to achieve significant improvements on node classification.

Future work will extend this method to more tasks demanding node representation learning with network structures.

Acknowledgements

This work is supported by CAS Project for Young Scientists in Basic Research (YSBR-040), National Natural Science Foundation of China (62373061), Beijing Natural Science Foundation (L232028), and the Project of ISCAS (ISCAS-JCMS-202401).

References

[Cai *et al.*, 2024] Jie Cai, Xin Wang, Haoyang Li, Ziwei Zhang, and Wenwu Zhu. Multimodal graph neural architecture search under distribution shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8227–8235, 2024.

[Cao *et al.*, 2022] Xianshuai Cao, Yuliang Shi, Jihu Wang, Han Yu, Xinjun Wang, and Zhongmin Yan. Cross-modal knowledge graph contrastive learning for machine learning method recommendation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM)*, pages 3694–3702. ACM, 2022.

[Chen and Zhang, 2020] Jiayi Chen and Aidong Zhang. HGMF: heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1295–1305, 2020.

[He *et al.*, 2024] Mingguo He, Zhewei Wei, Shikun Feng, Zhengjie Huang, Weibin Li, Yu Sun, and Dianhai Yu. Spectral heterogeneous graph convolutions via positive noncommutative polynomials. In *Proceedings of the ACM on Web Conference (WWW)*, pages 685–696. ACM, 2024.

[Hu *et al.*, 2019] Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4821–4830, 2019.

[Hu *et al.*, 2020] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *Proceedings of the Web Conference (WWW)*, pages 2704–2710, 2020.

[Huang *et al.*, 2024] Junjie Huang, Jiarui Qin, Yong Yu, and Weinan Zhang. Beyond graph convolution: Multimodal recommendation with topology-aware mlps. *CoRR*, abs/2412.11747, 2024.

[Jangra *et al.*, 2023] Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):296:1–296:36, 2023.

[Jia *et al.*, 2022] Xiangen Jia, Min Jiang, Yihong Dong, Feng Zhu, Haocai Lin, Yu Xin, and Huahui Chen. Multimodal heterogeneous graph attention network. *Neural Comput. Appl.*, 35(4):3357–3372, oct 2022.

[Kannan *et al.*, 2020] Amar Viswanathan Kannan, Dmitriy Fradkin, Ioannis Akrotirianakis, Tugba Kulahcioglu, Arquimedes Canedo, Aditi Roy, Shih-Yuan Yu, Malawade Arnav, and Mohammad Abdullah Al Faruque. Multimodal knowledge graph for deep learning papers and code. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*, pages 3417–3420, 2020.

[Kim *et al.*, 2023] Sein Kim, Namkyeong Lee, Junseok Lee, Dongmin Hyun, and Chanyoung Park. Heterogeneous graph learning for multi-modal medical data analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5141–5150, 2023.

[Li *et al.*, 2023a] Jiaqi Li, Guilin Qi, Chuanyi Zhang, Yongrui Chen, Yiming Tan, Chenlong Xia, and Ye Tian. Incorporating domain knowledge graph into multimodal movie genre classification with self-supervised attention and contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, pages 3337–3345, 2023.

[Li *et al.*, 2023b] Yilin Li, Jiaqi Zhu, Congcong Zhang, Yi Yang, Jiawen Zhang, Ying Qiao, and Hongan Wang. THGNN: an embedding-based model for anomaly detection in dynamic heterogeneous social networks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1368–1378. ACM, 2023.

[Liang *et al.*, 2022] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Fuchun Sun. Reasoning over different types of knowledge graphs: Static, temporal and multi-modal. 2022.

[Mo *et al.*, 2024] Yujie Mo, Feiping Nie, Ping Hu, Heng Tao Shen, Zheng Zhang, Xinchao Wang, and Xiaofeng Zhu. Self-supervised heterogeneous graph learning: a homophily and heterogeneity view. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[Ni *et al.*, 2019] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[Pezeshkpour *et al.*, 2018] Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing(EMNLP-IJCNLP)*, pages 3208–3218. Association for Computational Linguistics, October-November 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine*

Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research, 2021.

[Sun and Han, 2012] Yizhou Sun and Jiawei Han. *Mining Heterogeneous Information Networks: Principles and Methodologies*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers, 2012.

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

[Tao *et al.*, 2020] Zhulin Tao, Yinwei Wei, Xiang Wang, Xiangnan He, Xianglin Huang, and Tat-Seng Chua. Mgat: Multimodal graph attention network for recommendation. *Information Processing & Management*, 57(5):102277, 2020.

[Wang *et al.*, 2019] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032, 2019.

[Wang *et al.*, 2020] Jinguang Wang, Jun Hu, Shengsheng Qian, Quan Fang, and Changsheng Xu. Multimodal graph convolutional networks for high quality content recognition. *Neurocomputing*, 412:42–51, 2020.

[Wang *et al.*, 2023a] Hongjun Wang, Jiyuan Chen, Lun Du, Qiang Fu, Shi Han, and Xuan Song. Causal-based supervision of attention in graph neural network: A better and simpler choice towards powerful attention. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2315–2323. ijcai.org, 2023.

[Wang *et al.*, 2023b] Xin Wang, Benyuan Meng, Hong Chen, Yuan Meng, Ke Lv, and Wenwu Zhu. TIVA-KG: A multimodal knowledge graph with text, image, video and audio. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, pages 2391–2399. ACM, 2023.

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

[Wei *et al.*, 2023] Weiyi Wei, Jian Wang, Mengyu Xu, and Futong Zhang. Multimodal heterogeneous graph convolutional network for image recommendation. *Multimedia Systems*, pages 2747–2760, 2023.

[Wu *et al.*, 2024] Renjie Wu, Hu Wang, and Hsiang-Ting Chen. A comprehensive survey on deep multimodal learning with missing modality. *CoRR*, abs/2409.07825, 2024.

[Xu *et al.*, 2023] Baogui Xu, Chengjin Xu, and Bing Su. Cross-modal graph attention network for entity alignment. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, pages 3715–3723, 2023.

[Yang *et al.*, 2022] Yaming Yang, Ziyu Guan, Zhe Wang, Wei Zhao, Cai Xu, Weigang Lu, and Jianbin Huang. Self-supervised heterogeneous graph pre-training based on structural clustering. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 16962–16974, 2022.

[Yang *et al.*, 2023a] Jieyi Yang, Feng Zhu, Yihong Dong, and Jiangbo Qian. Fusing heterogeneous information for multi-modal attributed network embedding. *Applied Intelligence*, 53:1–20, 06 2023.

[Yang *et al.*, 2023b] Xiaocheng Yang, Mingyu Yan, Shirui Pan, Xiaochun Ye, and Dongrui Fan. Simple and efficient heterogeneous graph neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2023.

[Zhang *et al.*, 2019] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V. Chawla. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 793–803, 2019.

[Zhao *et al.*, 2024] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Comput. Surv.*, 56(9):216:1–216:36, 2024.

[Zhu *et al.*, 2024] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):715–735, 2024.

[Zhuo *et al.*, 2023] Jiaming Zhuo, Can Cui, Kun Fu, Bingxin Niu, Dongxiao He, Yuanfang Guo, Zhen Wang, Chuan Wang, Xiaochun Cao, and Liang Yang. Propagation is all you need: A new framework for representation learning and classifier training on graphs. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, pages 481–489. ACM, 2023.