HGMF: Heterogeneous Graph-based Fusion for Multimodal Data with Incompleteness

Jiayi Chen, Aidong Zhang
University of Virginia, Charlottesville, VA, USA
jc4td@virginia.edu,aidong@virginia.edu

ABSTRACT

With the advances in data collection techniques, large amounts of multimodal data collected from multiple sources are becoming available. Such multimodal data can provide complementary information that can reveal fundamental characteristics of real-world subjects. Thus, multimodal machine learning has become an active research area. Extensive works have been developed to exploit multimodal interactions and integrate multi-source information. However, multimodal data in the real world usually comes with missing modalities due to various reasons, such as sensor damage, data corruption, and human mistakes in recording. Effectively integrating and analyzing multimodal data with incompleteness remains a challenging problem. We propose a Heterogeneous Graph-based Multimodal Fusion (HGMF) approach to enable multimodal fusion of incomplete data within a heterogeneous graph structure. The proposed approach develops a unique strategy for learning on incomplete multimodal data without data deletion or data imputation. More specifically, we construct a heterogeneous hypernode graph to model the multimodal data having different combinations of missing modalities, and then we formulate a graph neural network based transductive learning framework to project the heterogeneous incomplete data onto a unified embedding space, and multi-modalities are fused along the way. The learning framework captures modality interactions from available data, and leverages the relationships between different incompleteness patterns. Our experimental results demonstrate that the proposed method outperforms existing graph-based as well as non-graph based baselines on three different datasets.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Classification and regression trees**; *Semi-supervised learning settings.*

KEYWORDS

multimodal fusion, data incompleteness, missing modalities, graph neural networks, heterogeneous graph

ACM Reference Format:
Jiayi Chen, Aidong Zhang. 2020. 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 '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3394486.3403182

KDD '20, August 23–27, 2020, Virtual Event, CA, USA

ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403182

1 INTRODUCTION

Multimodal data is dramatically increasing with the advances of data collection technologies. Data across multiple sources can provide complementary information that reveals the fundamental characteristics of real-world subjects and phenomenons [11]. Integrating multimodal data has promoted the performance in various application scenarios, such as object detection [7, 19], sentiment analysis [32, 33, 37], emotion recognition [13, 31, 33], and disease diagnosis [13]. Multimodal data fusion therefore has become a widely-studied topic in machine learning. Extensive prior works have been developed to combine modalities to learn joint representations or perform predictions, including traditional approaches, such as early fusion and late fusion [25, 34], and deep learning approaches, such as graph-based late fusion [6] and deep fusion [14, 31–33] which focuses on exploring multimodal interactions.

However, effectively integrating multimodal data with missing modalities remains a challenging problem. Missing modality is a common issue in real-world multimodal scenarios [5], and the missingness can be caused by various reasons such as sensor damage, data corruption, and human mistakes in recording. Missing modality imposes significant challenges to multimodal machine learning on incomplete data. There are mainly three technical challenges to be addressed. First, multimodal data with different combinations of missing modalities can have inconsistent dimensions and numbers of feature sets, and thus introduce difficulties to apply complete multimodal fusion models [6, 14, 31–34] that treat each independent multimodal instance in the same architecture. Second, effective multimodal fusion requires learning about complementary information, the modality-specific information as well as the multimodal interactions [11]. However, with the presence of missing modalities, relevant information cannot be directly derived from the incomplete individual data. Third, a large amount of missing data may dramatically reduce the size of data, resulting in the difficulty of learning high-dimensional interactive features from few samples.

Learning modality interactions and complementary information from incomplete multimodal data was less unexplored by previous multimodal machine learning research. Some previous works handle this problem using common strategies, such as deleting incomplete data samples or imputing missing modalities. Data deletion can dramatically reduce the number of training data and result in over-fitting of deep learning models, especially when a large amount of samples having different cases of missing data. Imputation based methods try to generate the missing modalities based on observed ones, using traditional imputation techniques such as

zero/average imputation and matrix completion, or deep learning based approaches [2, 18, 21, 36]. Then, multimodal fusion can be achieved simultaneously [20, 22, 26] or after imputation [2, 36]. However, imputation based methods may introduce extra noise to the original data which has negative impacts on the performances, and they are sometimes associated with complex auxiliary models such as deep generative models.

In this paper, we focus on handling the incomplete data without imputation. Several existing approaches [12, 29, 30] avoid using imputation to integrate modalities with incompleteness. Multi-source feature learning method [30] partitions the incomplete data to multiple complete subgroups, and then integrates representations of the subgroups as a sparse multi-task learning problem. Multi-hypergraph learning method [13] incorporates high-order relationships of subgroups and learn directly on the output. Despite these methods provide solutions, they ignore the complex intra- and inter-modal interactions and fail to learn the relationships among incomplete samples, and the reduction of available data will deteriorate the independent model performance. In this work, we formulate a new fundamental structure that facilitates the complex information extraction and integration from multimodal data with missing modalities, without data deletion or imputation.

The proposed method, namely **H**eterogeneous **G**raph-based **M**ultimodal **F**usion (HGMF), models the multimodal data with incompleteness in a heterogeneous graph structure, and then exploits a graph neural network-based transductive learning framework to extract complementary information from the highly interacted incomplete multi-modalities and fuse the information from different subspaces into a unified space. The proposed approach also tackles a series of technical challenges posed by the graph construction, the exploitation of multimodal interactions, and the integration of incomplete multimodal information in graph. In summary, the main contributions of this paper are as follows:

- We propose to model the highly interacted multimodal data with different incompleteness patterns in a heterogeneous hypernode graph (HHG). Such graph representation of data connects the instances that have various conditions of missing modalities, and helps to explore the complex multimodal interactions and data relationships.
- We propose a transductive learning framework based on graph neural network to perform the multimodal fusion of incomplete data within the constructed HHG. The key idea is to derive significant information from the instances having specific observations to learn those without them.
- We conduct experiments in multiple levels of data incompleteness to show that our method can deal with real scenarios with high percentage of missing data. We show the effectiveness of our model by comparing it with both inductive and transductive baselines on three datasets.

2 PROBLEM FORMULATION

In this section, we introduce our problem, task, and the key idea of the proposed method.

Incompleteness Patterns of Multimodal Data. For an M-modal dataset with incompleteness, there are $(2^M - 1)$ different combinations of missing modalities. In this paper, an *incompleteness pattern* refers to a combination of modalities. Therefore, an incomplete multimodal dataset has at most $(2^M - 1)$ incompleteness patterns. The block-wise structure shown in Figure 1 illustrates a trimodal ($M = 3$) dataset with seven incompleteness patterns [30]. The blocks with solid colors indicate the available modalities, and the others represent missing modalities. This figure also shows that instances can be divided into separate groups such that in each group all instances have the same incompleteness pattern and each instance only belongs to one pattern.

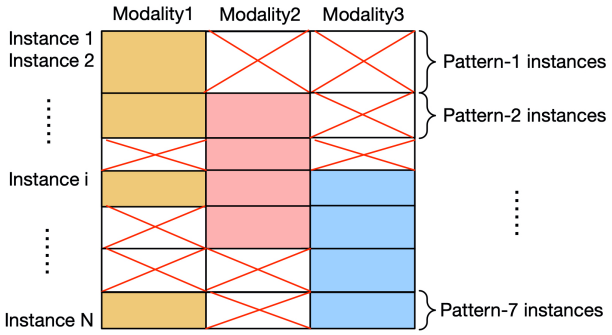

Figure 1: An illustration of a trimodal dataset ($M = 3$) with seven patterns of incompleteness.

Problem 2.1. **(Multimodal Fusion with Data Incompleteness).** Suppose M is the number of modalities (data sources), N is the number of samples, ψ is a function that maps each sample to a certain pattern, and $\phi(q) \subseteq \{1 \dots M\}$ indicates the set of available modalities for pattern q. Given a collection of incomplete multimodal data samples $\mathcal{D} = \{\widetilde{\mathbf{x}}_i\}_{i=1}^N$ as input, where each data sample consists of a set of available modalities $\widetilde{\mathbf{x}}_i = \{\mathbf{x}_{i,m}\}_{m \in \phi(\psi(i))}$, this paper aims to design a model that can capture modality interaction information from available data, and fuse multimodal data with different patterns using a same architecture. The learned joint representations are used for downstream tasks such as predictions.

Transductive Learning. In this paper, we formulate a transductive learning framework [13] to handle this problem without imputing missing data. Different from inductive learning, transductive learning (instance-based learning) directly incorporates the feature information implicit in other samples [38]. In this work, our key idea is that an incomplete data sample can derive the missing information from other samples having it within the transductive learning framework. Instances with different missing-data patterns can effectively exchange their modality-specific and interaction information, and multimodal fusion can be achieved along the way.

Among transductive learning variants, graph-based transductive learning methods achieved promising performance in practice [6, 38]. Recent advancements in graph neural networks (GNNs) also allow high-level features and high-dimensional representations to be learned from graph structural original data. Since graphs are powerful representations to model data relationships, in this paper, we use graphs to exchange significant information between multimodal instances, and formulate our problem in a novel GNN-based transductive learning framewok, HGMF.

3 METHODOLOGY

In this section, we present our HGMF method which is built based on a GNN transductive learning framework. The HGMF has three

stages: 1) modeling incomplete multimodal data in a proposed heterogeneous hypernode graph structure; 2) encoding the highly interacted multimodal data with the presence of missingness into more explicit modality-specific and cross-modal interaction information; and 3) aggregating and exchanging information among multimodal instances across different incompleteness patterns, through which all data can be fused into the same embedding space. Figure 2(a) illustrates the three-stage HGMF workflow using a simple four-data example. Note that in real scenarios, graphs can be larger and more complex than those shown in Figure 2(a). In the following, we will introduce the technical details in each stage.

3.1 Modeling Incomplete Multimodal Data with Heterogeneous Hypernode Graph

An incomplete multimodal dataset that has multiple missing-data patterns can be modeled as a k-NN affinity graph structure, where each node is an instance. However, as each instance contains multiple data sources, belongs to different incompleteness patterns, and has different feature spaces, we cannot use a simple affinity graph to model our problem.

To model multimodal data with incompleteness in a graph-based transductive learning framework, we first define a new family of graph structures, namely **H**eterogeneous **H**ypernode **G**raphs (HHG) whose structure and components are described below.

Definition 1 **(Heterogeneous Hypernode Graph).** A Heterogeneous Hypernode Graph (HHG) is denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \psi, \phi)$, containing the following components and properties.

- $\mathcal{V} = \{v_i\}_{i=1}^{N}$ is the *hypernode* set, where each hypernode is self-interacted. Different from simple graphs in which each node is associated with the same dimensional feature, hypernodes contains different numbers and dimensions of features, and each hypernode's features may be implicitly or explicitly interacted. In our problem, a hypernode refers to an multimodal instance, and we define the feature set of graph as $\mathcal{X} = \{\{\mathbf{x}_{i,m} | \forall m \in \phi(\psi(i))\}, 1 \leq i \leq N\}$.
- $\mathcal{E} = \{e_j\}_{j=1}^{|\mathcal{E}|}$ is the edge set. As we construct a k-NN affinity graph for instances, to more efficiently represent the high-order connections among nodes, we use the *hyperedges*[1] of hypergraphs instead of pairwise edges [6, 13]. A hyperedge is a subset of (hyper)nodes, connecting k instances who share some similar information, and showing a k-NN relationship among some nodes. Hyperedges $\mathcal{E}$ can be represented by an incidence matrix $\mathbf{H} \in \{0,1\}^{|\mathcal{V}| \times |\mathcal{E}|}$, where each row represents a hypernode v_i and each column represents a hyperedge e_j. For each entry in the incidence matrix, $\mathbf{H}(v_i, e_j) = 1$ indicates that the hypernode v_i is connected with some others through the hyperedge e_j. Each hyperedge e_j is associated with a single-valued weight w_j. In this paper, all edge weights equal to one.
- $\psi : \mathcal{V} \longmapsto \mathcal{T}$ defines the node pattern-mapping function. $\mathcal{T} = \{1, 2, ..., \overline{M}\}$ is the pattern set, where $\overline{M} = 2^M - 1$ is the number of incompleteness patterns observed from the dataset. The graph is heterogeneous because hypernodes (instances) have different incompleteness patterns, so we define ψ to distinguish multimodal instances.
- $\phi : \mathcal{T} \longmapsto \mathcal{P}(\mathcal{M}) \setminus \emptyset$ defines a function that maps a pattern to a combination of modalities, where $\mathcal{P}(\mathcal{M})$ denotes the power set (all subsets) of the set $\mathcal{M} = \{1, 2, ..., M\}$.

Heterogeneous Hypernode Graph Construction. Figure 2(b) shows an overview of the HHG graph construction process. On the left of the figure shows an trimodal incomplete dataset, where columns denote modalities and rows denote instances. Given such a multimodal dataset $\mathcal{D}$, one can easily obtain $\psi(\cdot), \phi(\cdot), \mathcal{V}$, and $\mathcal{X}$ based on data availability and corresponding features. On the right in Figure 2(b) is the seven-pattern HHG constructed from the input, where different colors denote different patterns. Note that each node here is a hypernode containing multi-modalities. We also provide an illustration of heterogeneous hypernodes in Figure 2(a)'s second subfigure, where each instance constructs a hypernode who contains at least one modality.

The bottom region in Figure 2(b) illustrates the hyperedge calculation process. $\mathcal{E}$ can be calculated from $\mathcal{D}$ as follows. As multimodal instances (hypernodes) have complex connections, each modality can provide a certain view of the data relationships. Motivated by [6, 13, 30], to capture the multi-view and high-order relationships among multimodal instances, we first reconstruct the blockwise incomplete dataset into B blocks according to different combinations of available modalities, and then calculate a set of hyperedges among all instances involved in each block. Let $\mathcal{V}_b$ and $\mathcal{M}_b$ denote the hypernodes and the modality set involved in the block b, respectively. We calculate the normalized distance between each pair of instances in a block ($\forall v_i, v_j \in \mathcal{V}_b$) as follows:

$$d^{(b)}(v_i, v_j) \triangleq \frac{1}{|\mathcal{M}_b|} \sqrt{\sum_{m \in \mathcal{M}_b} ||u_m(\mathbf{x}_{i,m}) - u_m(\mathbf{x}_{j,m})||_2^2 / Z_m}, \quad (1)$$

where $Z_m = \sum_{i,j \in \mathcal{V}_b} ||u_m(\mathbf{x}_{i,m}) - u_m(\mathbf{x}_{j,m})||_2^2$ and $\{u_m(\cdot), m = 1...M\}$ are the pre-trained unimodal representation learning models that are used to initialise shallow unimodal features. After calculating all distances, each hyperedge is calculated using k nearest neighbor method centered with each node [13]. As shown at the bottom region in Figure 2(b), for all pre-defined blocks, B sets of hyperedges are calculated independently depending on the feature extraction models' parameters. Suppose their incidence matrix are $\{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_B\}$, the final incidence matrix for HHG is the concatenation form $\mathbf{H} = [\mathbf{H}_1; \mathbf{H}_2; ...; \mathbf{H}_B]$. In this way, instances that do not have certain modalities can be connected with those that have the modalities, and then the incomplete data problem can be alleviated. In Figure 2(a)'s second subfigure, the hyperedges in different colors connects instances according to different blocks.

3.2 Intra-hypernode Encoder

After constructing the input graph $\mathcal{G}$ and feature set $\mathcal{X}$, we propose an intra-hypernode encoder to capture complementary information [11] from the highly interacted modalities with the presence of missing data. The intra-hypernode encoder, whose architecture is shown in Figure 2(c), consists of two components: 1) *Unimodal Embedding Networks* take unimodal data as input, and output unimodal embeddings, and 2) *Feature Interaction Networks* captures the modality interactions among these embeddings, and extract

[1] hyperedges: edges in hypergraphs are called hyperedges or hyperlinks. A hyperdge connects two or more than two vertices.

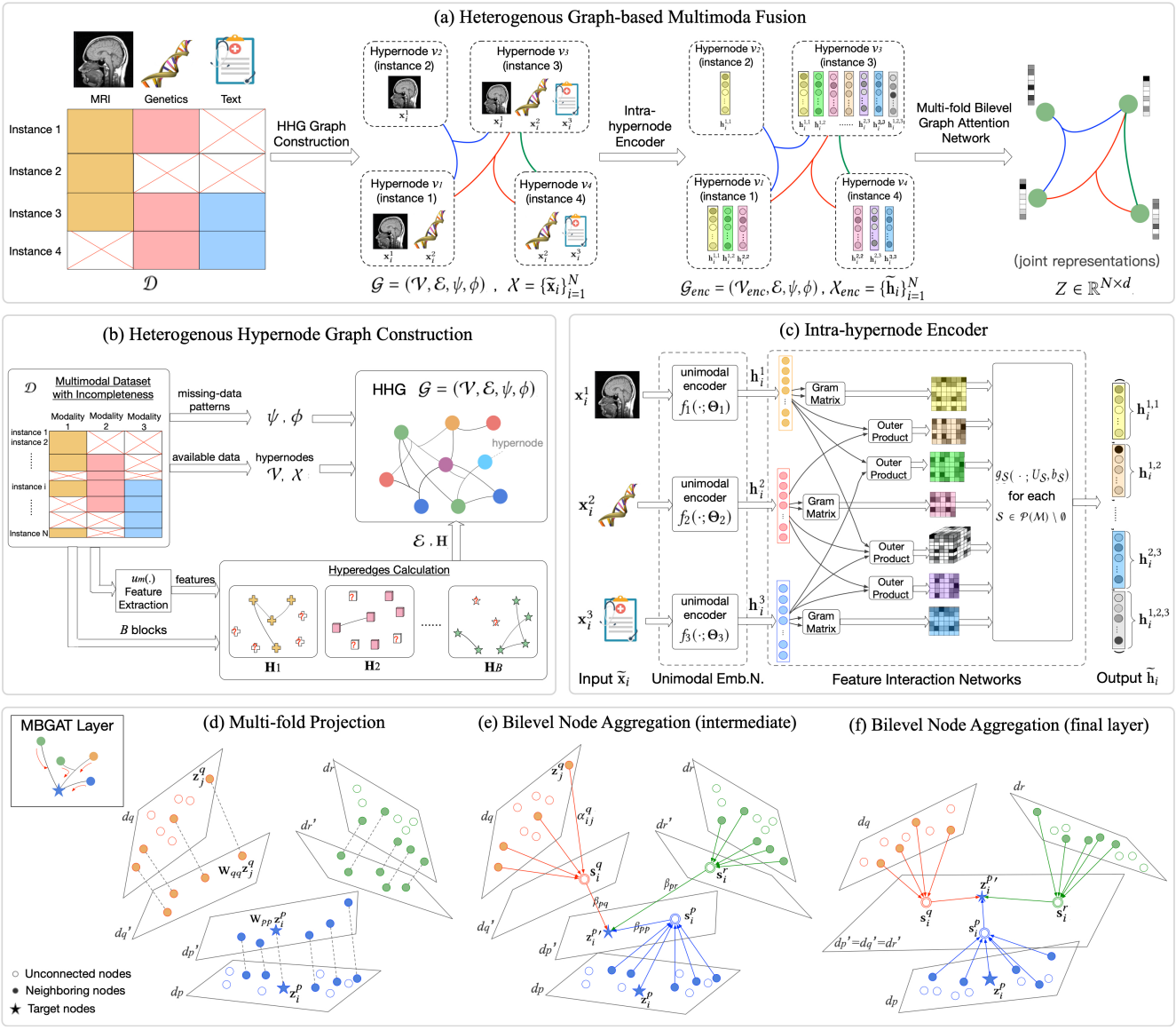

Figure 2: Overview of Heterogeneous Graph-based Multimodal Fusion (HGMF) with trimodal setting ($M = 3$). (a) Illustration of the three-stage workflow of HGMF. It shows a four-data example, but in real tasks, the graph can be larger in scale and more complex. From left to right, $\mathcal{D}$ include four instances with four patterns of missing modalities; $\mathcal{G}$ and $\mathcal{X}$ are constructed after HHG graph construction stage; $\mathcal{G}_{enc}$ and $\mathcal{X}_{enc}$ are obtained through intra-hypernode encoding stage; and, multimodal instances are finally fused into joint representations Z through multiple MBGAT layers' graph embedding stage. (b) Pipeline of graph construction. (c) Architecture of intra-hypernode encoder, which takes as input each hypernode feature set $\widetilde{\mathbf{x}}_i$ and output new feature set $\widetilde{\mathbf{h}}_i$. Note that for incomplete instances, partial related neurons are blocked based on which modalities are missing. (d) Illustration of the multi-fold embedding space projection at an MBGAT layer. (e) Illustration of the bilevel node aggregation at any intermediate MGBAT layer. (f) Illustration of the bilevel node aggregation at the final MGBAT layer.

complementary information (modality-specific and cross-modal interaction information) from them.

3.2.1 ***Unimodal Embedding Networks.*** Since the original data in $\mathcal{X}$ is very high-dimensional, sparse, and inconsistent with respect to data structure, it is hard to calculate interactions among original modalities. We thus setup a series of *source-specific* Deep Neural Networks (DNNs) to learn compressed and rich feature representations from unimodal original data.

Based on real data structures in the datasets on which we perform our models, we mainly consider three types of architectures to build unimodal embedding networks: 1) Convolutional Neural Networks (CNN) for embedding image modalities; 2) Bidirectional Long-short Term Memory (BI-LSTM) for embedding sequential modalities, such as video, free texts (e.g., clinical records) and spoken language; and 3) Stacked fully connected layers followed by nonlinear activation

functions, for embedding high-dimensional or sparse feature-based modalities, such as gene expressions.

Suppose $f_m(\cdot;\Theta_m)$ be the m's unimodal embedding network with learnable parameter Θ_m. For each hypernode v_i whose content is $\widetilde{\mathbf{x}_i} = \{\mathbf{x}_{i,m}\}_{m\in\phi(\psi(i))}$, its modality-$m$ is embedded as

$$\mathbf{h}_i^m = f_m(\mathbf{x}_{i,m};\Theta_m), \tag{2}$$

where $\mathbf{h}_i^m \in \mathbb{R}^{F_m}$. F_m is the embedding dimension of modality-m.

3.2.2 ***Feature Interaction Networks.*** A hypernode contains unimodal components that are highly interacted. Such modality interactions are high-order and implicit, and therefore difficult to be represented. The goal of feature interaction networks is to capture such interactions, to extract modality-specific and cross-modal interaction information from them.

The high-order modality interactions can appear individually, between each pair of modality and among more than two modalities. Let $\mathcal{P}(\cdot)$ denote the power set operation and $\mathcal{M} = \{1, 2, ..., M\}$. As each subset $\forall\mathcal{S} \in \mathcal{P}(\mathcal{M}) \setminus \emptyset$ denotes a combination of modalities, we can learn from each $\mathcal{S}$ one type of multimodal interaction and a piece of information, namely factor.

A hypernode's complementary information consists of many factors. Let each factor be represented by a F'-dimensional vector. A factor can be calculated as follows.

If there is only one element in $\mathcal{S}$ (i.e., $|\mathcal{S}| = 1$), meaning that we are calculating *modality-specific information* for the modality $m \in \mathcal{S}$, we can calculate modality-specific information $\mathbf{h}_i^{m,m}$ as

$$\begin{aligned}\overline{\mathbf{h}}_i^m &= g_m(\mathbf{h}_i^m;\mathbf{U}_m,\mathbf{b}_m)\\ \mathbf{G}_i^m &= (\mathbf{h}_i^m)(\mathbf{h}_i^m)^T\\ \mathbf{h}_i^{m,m} &= g_{m,m}(\mathbf{G}_i^m;\mathbf{U}_{m,m},\mathbf{b}_{m,m}) + \overline{\mathbf{h}}_i^m,\end{aligned} \tag{3}$$

where $\mathbf{U}_m \in \mathbb{R}^{F'\times F_m}$, $\mathbf{b}_m \in \mathbb{R}^{F'}$, $\mathbf{U}_{m,m} \in \mathbb{R}^{F'\times(F_m)^2}$ and $\mathbf{b}_{m,m} \in \mathbb{R}^{F'}$ are parameters of the neural networks $g_m(\cdot)$ and $g_{m,m}(\cdot)$, respectively. $\mathbf{G}_i^m \in \mathbb{R}^{F_m\times F_m}$ is the Gram matrix of unimodal embedding $\mathbf{h}_i^m$, which represents the covariance, the feature self-interaction information; and $\overline{\mathbf{h}}_i^m$ can be viewed as the mean, low-dimensional, and specific information for modality-m.

If there are more than one elements in $\mathcal{S}$ (i.e., $|\mathcal{S}| > 1|$), meaning that we are calculating *cross-modal interaction information* among all unimodal embeddings $\{\mathbf{h}_i^m|\forall m \in \mathcal{S}\}$. Inspired by [31], we can calculate cross-modal interaction information $\mathbf{h}_i^{\mathcal{S}}$ as

$$\begin{aligned}\mathbf{C}_i^{\mathcal{S}} &= \otimes_{m\in\mathcal{S}}\mathbf{h}_i^m\\ \mathbf{h}_i^{\mathcal{S}} &= g_{\mathcal{S}}(\mathbf{C}_i^{\mathcal{S}};\mathbf{U}_{\mathcal{S}},\mathbf{b}_{\mathcal{S}}),\end{aligned} \tag{4}$$

where $\mathbf{C}_i^{\mathcal{S}}$ represents the $|\mathcal{S}|$-fold cross-product of the involved unimodal embeddings, and $\mathbf{U}_{\mathcal{S}} \in \mathbb{R}^{F'\times(\prod_{m\in\mathcal{S}} F_m)}$ and $\mathbf{b}_{\mathcal{S}} \in \mathbb{R}^{F'}$ is the learnable weights of the neural network $g_{\mathcal{S}}(\cdot)$. A special case is that, for example, the bimodal interaction information between m and k can be extracted as $\mathbf{h}_i^{m,k} = g_{m,k}((\mathbf{h}_i^m)(\mathbf{h}_i^k)^T;\mathbf{U}_{m,k},\mathbf{b}_{m,k})$.

3.2.3 ***Summary.*** In this section, our intra-hypernode encoder leverages all combinations of unimodal-specific and cross-modal interactions, and extracts pieces of complementary information from multi-modalities with the presence of missing data.

The architecture shown in Figure 2(c) is shared by all hypernodes. Intra-hypernode encoder takes as input each hypernode feature set $\widetilde{\mathbf{x}_i} = \{\mathbf{x}_{i,m}\}_{m\in\phi(\psi(i))}$ and output new feature set $\widetilde{\mathbf{h}_i} = \{\mathbf{h}_i^{\mathcal{S}}\}_{\mathcal{S}\in\mathcal{P}(\phi(\psi(i)))\setminus\emptyset}$. Then, we obtain a new heterogeneous hypernode graph $\mathcal{G}_{enc} = (\mathcal{V}_{enc}, \mathcal{E}, \psi, \phi)$ associated with a new feature set $\mathcal{X}_{enc} = \{\widetilde{\mathbf{h}_i}\}_{i=1}^N$, which is illustrated in Figure 2(a).

3.3 Multi-fold bilevel Graph Attentional Layer

In the previous step, the intra-hypernode encoder outputs new HHG $\mathcal{G}_{enc}$ and feature set $\mathcal{X}_{enc}$, which contains more explicit modality-specific and cross-modal interaction information than the input feature set. The hypernodes in $\mathcal{G}_{enc}$ are also heterogeneous, because hypernodes with different incompleteness patterns contain different numbers and categories of factors; there are a total of $(2^{|\phi(\psi(i))|} - 1)$ factors calculated for hypernode v_i.

In this section, we focus on learning the interactions between different incompeteness patterns, and propose to simultaneously fuse multimodal data with different missingness within a graph-based transductive learning architecture. Specifically, we propose to solve a sub-problem described as follows.

Problem 3.1. **($\overline{M}$-fold Heterogeneous Graph Embedding)**. Given the heterogeneous hypernode graph $\mathcal{G}_{enc} = (\mathcal{V}_{enc}, \mathcal{E}, \psi, \phi)$, where the node set can be divided into $\overline{M} = |\mathcal{T}|$ non-overlapping subsets $\mathcal{V}_{enc} = \{\mathcal{V}_p|\forall p \in \mathcal{T}\}$ based on $\psi(\cdot)$, and each node $v_i \in \mathcal{V}_p$ is associated with a set of F'-dimensional vectors $\widetilde{\mathbf{h}_i} = \{\mathbf{h}_i^{\mathcal{S}}|\forall\mathcal{S} \in \mathcal{P}(\psi(p)) \setminus \emptyset\}$, the task is to learn to map the heterogeneous hypernodes in $\overline{M}$ embedding spaces, into a homogeneous embedding space $\mathbf{Z} \in \mathbb{R}^{N\times d}$.

By solving this sub-problem, hypernodes with different incompleteness patterns can be projected onto the same feature space; multimodal instances with missing information (missing factors) can derive such information from others; and, the final node embeddings can be the fused multimodal representations.

To solve this sub-problem, in this section, we propose **M**ulti-fold **B**ilevel **G**raph **A**ttention Ne**t**works (MBGAT) inspired by graph attention networks (GATs) [24]. In the following, we state the overall goal of each MBGAT layer, and then introduce technical details.

3.3.1 ***Overview.*** At each MBGAT layer, the goal is to project the existing features in $\overline{M}$ spaces onto $\overline{M}$ new spaces (see Figure 2(d-f)) that are close to each other. However, aggregating information from heterogeneous nodes is challenging as the relationships between different feature spaces are unknown. In the context of multimodal fusion, such difficulty comes from the unknown relationships between different incompleteness patterns.

To tackle this problem, inspired by self-attention mechanism [23] and GATs [24], we design a bilevel attention strategy to aggregate neighborhood information among different patterns. An MBGAT layer consists of two components: *multi-fold intra-pattern aggregation* that aggregates nodes in the same space independently, and *inter-pattern aggregation* that learns pattern relationships, and fuses all neighbors into one target space.

At each MBGAT layer, we represent the multi-space inputs as $\mathbf{z} = \{\{\mathbf{z}_i^p|\forall v_i \in \mathcal{V}_p\}|\forall p \in \mathcal{T}\}$, where $\mathbf{z}_i^p \in \mathbb{R}^{d_p}$ is the d_p-dimensional feature associated with the pattern-p node v_i. The layer's multi-space outputs are represented as $\mathbf{z}' = \{\{\mathbf{z}_i^{p\prime}|\forall v_i \in \mathcal{V}_p\}|\forall p \in \mathcal{T}\}$, $\mathbf{z}_i^{p\prime} \in \mathbb{R}^{d'_p}$, where d'_p is the dimension of the new feature space of

pattern-p. Note that at the first layer, we initialize the input node featues $\mathbf{z}^{(0)}$ by concatenating all feature vectors in hypernodes, since the previously extracted features within a hypernode are separate pieces of information. In the following, we present the technical details on how to aggregate neighborhood information for a target node v_i whose pattern is $p = \psi(i)$.

3.3.2 ***Multi-fold Intra-pattern Aggregation.*** As the lower-level aggregation, we focus on aggregating neighbors in the same feature space (multimodal instances that miss the same modalities).

Multi-fold Projection. Each node should be projected onto its new and lower-dimensional feature space to get prepared for aggregation. As we prepare all nodes at the same time, each node in any pattern's feature space will need to be combined with nodes in any other pattern's space. We therefore define $\{\mathbf{W}_{pq}|\forall p, q \in \mathcal{T}\}$ as our $|\mathcal{T}|$-fold projection scheme, where $\mathbf{W}_{pq} \in \mathbb{R}^{d'_q \times d_p}$ is learnable matrix that projects nodes from the pattern-p's feature space to the pattern-q's *new* feature space. Figure 2(d) illustrates the multi-fold projection, where nodes in different colors (different patterns) are projected onto different spaces.

Intra-patten Aggregation. Suppose $\mathcal{N}_q(i)$ denote the pattern-q neighboring node set of v_i, which can be defined as $\mathcal{N}_q(i) = \{v_j|\forall v_j \in \mathcal{V}_q \wedge (\mathbf{H}\mathbf{H}^T)_{ij} > 0\}$, where $\mathbf{H}$ is the incidence matrix (constructed in Section 3.1). We then calculate the importance of each neighboring node in $\mathcal{N}_q(i)$ to the target node v_i by performing the attention mechanism, $\vec{\mathbf{a}}_q \in \mathbb{R}^{2d'_q \times 1}$. The calculated attention coefficients for each pattern-q neighbor v_j is

$$\alpha^q_{ij} = \frac{exp(LeakyReLU(\vec{\mathbf{a}}_q[\mathbf{W}_{pq}\mathbf{z}^p_i; \mathbf{W}_{qq}\mathbf{z}^q_j]))}{\sum_{k\in\mathcal{N}_q(i)} exp(LeakyReLU(\vec{\mathbf{a}}_q[\mathbf{W}_{pq}\mathbf{z}^p_i; \mathbf{W}_{qq}\mathbf{z}^q_k]))}. \quad (5)$$

Finally, the intra-pattern aggregation result from all pattern-q neighbors of node v_i is

$$\mathbf{s}^q_i = \sigma\left(\sum_{j\in\mathcal{N}_q(i)} \alpha^q_{ij}\mathbf{W}_{qq}\mathbf{z}^q_j\right) \quad (6)$$

where $\sigma(\cdot)$ denotes the sigmoid function. In Figure 2(e) and (f), we can see that nodes in the same colors (same pattern) are aggregated to a certain double-circled feature points on the new space.

3.3.3 ***Inter-pattern Aggregation.*** After aggregating the neighborhood information within each pattern, we aim to learn the relationships between different patterns, so that multimodal instances that have different missing modalities can derive information from each other. To achieve the goal, we perform inter-pattern aggregation as the higher-level aggregation.

Similarly, given the intra-pattern aggregation results $\{\mathbf{s}^1_i, \mathbf{s}^2_i, ..., \mathbf{s}^{\overline{M}}_i\}$, $\mathbf{s}^q_i \in \mathbb{R}^{d'_q}$, we can calculate the importance of the pattern-q neighbors to the pattern-p target by performing the attention mechanism, $\mathbf{b}_p \in \mathbb{R}^{2d'_p \times 1}$. The calculated attention coefficients is:

$$\beta_{pq} = \frac{exp(LeakyReLU(\vec{\mathbf{b}}_p[\mathbf{V}_{pp}\mathbf{s}^p_i; \mathbf{V}_{qp}\mathbf{s}^q_j]))}{\sum_{r\in\mathcal{T}} exp(LeakyReLU(\vec{\mathbf{b}}_p[\mathbf{V}_{pp}\mathbf{s}^q_i; \mathbf{V}_{rp}\mathbf{s}^r_i]))}, \quad (7)$$

where $\mathbf{V}_{qp} \in \mathbb{R}^{d'_p \times d'_q}$ denotes the space-to-space transformation from pattern-q to pattern-p. Once obtained, the attention coefficients are used to compute a linear combination of intra-pattern aggregation results. Finally, we can update the embedding for target node v_i as:

$$\mathbf{z}^{p\prime}_i = \sigma\left(\sum_{q\in\mathcal{T}} \beta_{pq}\mathbf{V}_{qp}\mathbf{s}^q_i\right). \quad (8)$$

As shown in Figure 2 (e) and (f), double-circled points (intra-aggregation results) are aggregated to the blue star node on d'_p space (the node v_i's new embedding point on pattern-p's new space).

Algorithm 1 HGMF

Input data: $\mathcal{M}, \mathcal{T}, \mathcal{V}, \phi, \psi, \mathcal{D} = \{\widetilde{\mathbf{x}}_i\}_{i=1}^N, \mathcal{Y}_{trn}$
Input parameters: k, η
Initialise networks with ramdom parameters θ_e, θ_g, θ_p.
$\mathbf{H} \leftarrow kNN(\mathcal{D}, k)$ using Eq. (1)
$\mathcal{X} = \{\{\mathbf{x}_{i,m}\}_{m\in\phi(\psi(i))\subseteq\mathcal{M}}\}_{i=1}^N \leftarrow \mathcal{D}$
$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \psi, \phi) \leftarrow \mathbf{H}, \mathcal{X}$
$t \leftarrow 0$
while stopping condition is not met **do**
  for $v_i \in \mathcal{V}$ **do**
    $\widetilde{\mathbf{h}}_i \leftarrow \text{InHyNEnc}_{\theta_e}(\widetilde{\mathbf{x}}_i)$ using Eq. (2-4)
  end for
  $\mathcal{G}_{enc}, \mathcal{X}_{enc} \leftarrow \{\widetilde{\mathbf{h}}_i\}_{v_i\in\mathcal{V}}$
  $\mathbf{Z} \leftarrow \text{MBGAT}_{\theta_g}(\mathcal{G}_{enc}, \mathcal{X}_{enc})$ using Eq. (5-8)
  $\hat{\mathbf{Y}} \leftarrow \text{Prediction}_{\theta_p}(\mathbf{Z})$
  $\mathcal{L} \leftarrow \hat{\mathbf{Y}}_{trn}, \mathcal{Y}_{trn}$ using Eq. (9)
  Update $\theta_e, \theta_g, \theta_p \leftarrow \min_{\theta_e,\theta_g,\theta_p}\mathcal{L}$
  $t \leftarrow t + 1$
end while
return: $\theta^*_e, \theta^*_g, \theta^*_p$, and $\hat{\mathbf{Y}}_{test}$

3.3.4 ***Summary.*** In this section, we proposed MBGAT, which performs multimodal fusion through an $\overline{M}$-fold heterogeneous graph embedding procedure. At each MBGAT layer, the two levels of aggregation enable each node to receive information from its $\overline{M}$ patterns of neighboring nodes, in which learned attention coefficients are responsible to handle how different modality interaction information exchanges between incomplete data.

We stack multiple MBGAT layers, so that the heterogeneous multimodal nodes can be embedded and fused layer by layer. In this paper, we let $L = 2$ in all experiments. Note that at the final layer, all patterns of nodes are projected to a consistent feature space (see Figure 2 (f)). In other words, we let $d = d^{(L)}_1 = d^{(L)}_2 = d^{(L)}_3 = ... = d^{(L)}_{\overline{M}}$. Finally, the output embedding $\mathbf{Z} = \mathbf{Z}^{(L)} \in \mathbb{R}^{N\times d}$ is the fused representations for all multimodal instances.

4 EXPERIMENTS

We conduct experiments with the aim of answering two questions: 1) How does HGMF perform for multimodal classification tasks with different percentages of missingness? And, 2) How does HGMF perform compared with inductive and transductive baselines?

4.1 Data

4.1.1 ***Datasets.*** We perform experiments in both bimodal and trimodal settings, considering three datasets. 1) **ModelNet40** [28] is

Table 1: Statistics of Datasets (NoIS: number of incompletness scenarios created from the original dataset; M: number of modalities; and $|C|$: number of classes)

Dataset	Train/ Valid/Test	M	NoIS	$\|C\|$
ModelNet40	7,387/1,231/3,693	2	4	40
NTU	1,207/201/604	2	4	67
IEMOCAP	2,680/447/1,340	3	4	2

a large scale 3D CAD dataset, containing 12,311 3D shapes covering 40 common categories, including airplane, bathtub, bed, bench, bookshelf, bottle, bowl, cone, cup, and so on. 2) **NTU** [3] dataset is composed of 2,012 3D shapes from 67 categories, including car, chair, chess, chip, clock, cup, pen, plant leaf and so on. 3) **IEMOCAP** [1] dataset consists of a collection of 151 videos of recorded dialogues, with 2 speakers per session for a total of 302 videos across the dataset. Each segment is annotated for the presence of 9 emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral). ModelNet40 and NTU are used as bimodal datasets, and IEMOCAP is used as trimodal dataset. See Appendix A.1 for more descriptions about the datasets and data sources.

4.1.2 ***Incompleteness of Datasets.*** We evaluate the performance of HGMF under different percentages of data incompleteness. From each multimodal dataset, we prepare the input data by creating several blockwise incomplete multimodal scenarios. Given a Multimodal Incompleteness Ratio (MIR) ρ% and suppose the dataset is M-modal, we randomly delete data from the original complete datasets such that a total of ρ% instances have different conditions of missing modalities. In particular, for each incomplete scenario, we let each incompleteness pattern has $N \times \rho/(2^M - 1)$% instances. For example, given a bimodal dataset, for each class, we randomly sample $N \times \rho/2$% instances to remove their first modality, and sample $N \times \rho/2$% from the rest to remove the second modality.

4.1.3 ***Data Split.*** All datasets are split into training, validation, and testing sets. In general, to ensure balanced datasets, for each class and each incompleteness pattern, 60% data are used for training, 10% for validation, and 30% for testing. Table 1 shows a summary of the datasets and data split information.

4.2 Baseline Models

To achieve a comprehensive and comparative analysis of HGMF, we compare it with previously proposed neural multimodal fusion models, which can be divided into two categories: 1) inductive multimodal fusion models, including **Concat**, **Tensor Fusion Network** (TFN) [31], **Low-rank Fusion Network** (LFM) [14], and **Multitask Multimodal Learning** (MTL); and 2) the transductive model **Hypergraph Neural Network** (HGNN) [6]. See Appendix A.2 and A.3 for more details of baselines and reproducibility.

4.3 Experimental Setup

4.3.1 ***Model settings.*** We employ Pytorch3 to implement all baselines and the proposed HGMF with both bimodal and trimodal settings. See Appendix A.4 for more details about model settings.

4.3.2 ***Model Training.*** The overall training procedure is in Algorithm 1. Since we construct multimdoal instances in an HHG graph structure, we formulate the training of our data fusion system HGMF as a semi-supervised node classification task [10, 24]. After embedding the hypernodes in Section 3.3, the fused representations of incomplete multimodal instances is $Z^{(L)} \in \mathbb{R}^{N \times d}$. For $|C|$-class classification tasks, given the set of labels for training data $\mathcal{Y}_{trn} = \{y_i | \forall v_i \in \mathcal{V}_{trn} \subset \mathcal{V}\}$ where $y_i \in \{c_1, c_2, ..., c_{|C|}\}$, we minimize the cross-entropy loss defined as follows:

$$\mathcal{L} = - \sum_{i \in \mathcal{V}_{trn}} \sum_{c=1}^{|C|} y_i \cdot \log(\text{softmax}(p(\mathbf{z}_i^{(L)}; \theta_p))). \tag{9}$$

where θ_p is the parameter of the classifier $p(\cdot)$, a fully connected deep neural network shared by all nodes' fused representation.

Optimization. Parameters of intra-hypernode encoder and multi-fold bilevel graph attention network are initialized with uniform distribution. Before training the entire network, we do not consider node connections, treating each node independently to pre-train the intra-hypernode encoder. Then, we train the whole model parameters via the Adam optimizer [9] with tuned learning rates. We repeat the training iterations until the validation set's accuracy change between two consecutive iterations is sufficiently small.

4.4 Results and Analysis

We perform classification tasks to evaluate our model against baselines. For multi-class datasets, we report classification accuracy Acc-K where K denotes the number of classes. For binary classification datasets, we report F1 score. The results of our comparative evaluation experiments are summarized in Table 2 and Table 3. We examine the efficacy of the proposed HGMF model on both complete scenarios and incomplete scenarios.

4.4.1 ***Comparison on complete data.*** We first evaluate the effectiveness of the proposed method on complete multimodal data. In such scenarios, there is only one pattern in the dataset.

In Table 2, the results (columns CPL) on both datasets are higher than baselines. It proves that the proposed model can also be used in complete multimoda fusion. Compared with inductive learning, our improvement on NTU dataset is higher than that on ModelNet40 dataset. Our model performs only slightly better than GNN-based methods because the two modalities that are also used in HGNN may not highly-interacted.

In Table 3, we only compare with non-graph based inductive learning methods, because the graph-based baseline HGNN cannot deal with modalities in different dimensional tensor in IEMOCAP dataset, and also cannot handle the modality interactions. From the comparisons with inductive learning, our model's results outperform Concat and MTL, and either higher or similar to TMF and LMF. It is because baselines do not need to impute data in complete scenarios, so that our model and baselines are under the same circumstances.

4.4.2 ***Comparison on incomplete data.*** Now we consider the more realistic setting where blockwise missing modalities is present. We evaluate the influences of missing modalities by changing the multimodal incompleteness ratio from 30% to 75% with a intermittent 15%. As shown in both Table 2 and Table 3, our method tends to outperform baselines while there are more missing modalities.

Table 2: Test Accuracy (%) on ModelNet40 and NTU (M=2) datasets compared with baselines with various MIRs (CPL: complete).

Method	ModelNet40					NTU				
	CPL	30%	45%	60%	75%	CPL	30%	45%	60%	75%
$\text{Concat}_{imput\_zero}$	96.54	95.72	94.89	92.33	91.41	89.34	88.45	86.65	83.67	81.56
TFN_{imput_zero}[31]	98.16	96.02	95.48	93.81	93.3	**93.03**	**91.40**	87.97	85.07	84.72
LMF_{imput_zero}[14]	97.62	96.10	95.94	93.30	92.47	90.94	89.73	85.62	82.25	78.72
HGNN_{imput_zero}[6]	97.80	96.52	**96.10**	93.80	91.83	92.73	91.13	87.58	85.41	84.24
MTL	97.40	96.13	95.12	94.12	93.2	90.70	89.91	86.36	83.08	82.39
HGMF (ours)	**98.29**	**97.20**	96.02	**94.78**	**93.87**	**92.38**	91.22	**88.77**	**86.41**	**85.89**

Table 3: F1 scores on three emotion categories in IEMOCAP dataset ($M = 3$) compared with baselines in different incompleteness scenarios (EMO&IS: emotion categories and incompleteness scenarios; CPL: complete).

EMO&IS		Concat	TFN	LMF	MTL	**HGMF** (ours)
Happy	CPL	86.26	88.72	**89.69**	87.52	88.87
	30%	86.54	**88.74**	88.56	87.43	88.70
	45%	85.92	87.51	86.38	86.21	**87.93**
	60%	85.73	87.00	85.87	86.74	**87.24**
	75%	83.27	86.53	84.56	86.02	**86.02**
Sad	CPL	83.69	85.09	85.45	84.71	**85.72**
	30%	83.23	**85.25**	84.35	83.97	84.67
	45%	82.61	**84.58**	83.47	83.88	84.55
	60%	81.35	82.04	81.26	81.97	**83.33**
	75%	80.69	81.95	79.89	81.06	**82.32**
Angry	CPL	86.74	88.22	**88.74**	87.75	88.38
	30%	85.93	88.02	87.59	87.66	**88.14**
	45%	84.86	87.42	86.38	86. 03	**87.81**
	60%	83.29	86.17	85.25	85.26	**87.34**
	75%	83.71	85.46	84.68	85.02	**86.89**

From Tables 2 and 3, as more modalities are missing, the performances of Concat and LMF drop dramatically, while TFN, MTL and the proposed HGMF do not drop too much. It is because Concat and LMF do not explore much inter-modalitiy interactions, and their network neurons can be significantly affected by attacked values at the beginning. Also, when the missing ratio is not lower enough (i.e., less than 45%), the results show that TFN does not drop too much as the proposed method. It may be due to that the zero imputation can be viewed as dropout layer at the beginning, and the dropout at a low rate does not influence the higher-level neurons too much.

5 RELATED WORK

Our work is relevant to three lines of work: 1) deep multimodal fusion for complete data, 2) multimodal data analysis for incomplete data, and 3) graph-based transductive learning.

Complete Multimodal Fusion. The majority of prior studies on deep multimodal fusion assume complete feature sets. Early fusion methods refer to concatenating multimodal data at the input level [16, 17], while late fusion methods [25, 34] integrate unimodal outputs. Graph-based methods such as hypergraph neural networks (HGNN) [6] perform early fusion (concatenation) as well as late fusion which exploits graph structural relationships among unimodal representations to integrate outputs. However, these methods have limited capabilities in exploring complementary information from high-order modality interactions, and cannot deal with missing uni-modalities. Recent methods that perform intermediate fusion include multimodal sequential learning [32, 33] for sequential data (time series, language, audio and video), and post-dynamic learning for general data [14, 31]. However, these works cannot model multimodal interactions with the presence of missing modalities.

Incomplete Multimodal Data Analysis. Imputation methods [2, 22, 36] that complete or generate missing modalities may introduce extra noise to the fusion process. Non-imputation methods such as multi-source learning [30] and multi-hypergraph learning [13] first partition the incomplete data to multiple complete subgroups, and then integrate subgroup representations using multi-task learning or shallow graph learning through graph Laplacian. However, these works fail to effectively model the interactions between modalities with missingness, and fail to explore the relationships between different incompleteness patterns.

Graph-based Transductive Learning. Several graph-based transductive learning models designed generally are also related to our work. Graph Attention Networks (GATs) [24] compute attention coefficients using an edge-wise mechanism, which is extended in our work to learn a heterogeneous hypernode graph where there is not an immediately-obvious structure. We employ the graph attention mechanism to learn the unknown relationships between different incompleteness patterns within a heterogeneous hypernode graph. In addition, several GNN variants [27, 35] propose to handle node embedding in heterogeneous graphs. Recent approaches, such as HGNN [6] and multi-hypergraph learning [13], perform late fusion on graphs constructed from complete or incomplete multimodal data, through traditional graph Laplacian or graph neural networks. However, these methods fail to learn the high-order relationships between data with different missingness patterns. However, the above heterogeneous graph and hypergraph learning methods cannot deal with the more complex multimodal hypernode structure in our problem.

6 CONCLUSION

We have presented the heterogeneous graph-based multimodal fusion (HGMF) framework, a novel multimodal fusion method that exploit a heterogeneous hypernode graph (HHG) structure to capture modality interactions from incomplete modalities (intra-hypernode

encoder) as well as learn the relationships among different incompleteness patterns (MBGAT). The idea is to exploit the powerful graphs representations to enable incomplete data samples to derive relevant missing information from other samples who have such information. Through the information integration within HHG, the proposed HGMF framework effectively fuses multimodal data into joint representations and makes decisions based on them. Our experimental results demonstrated the significance of our approach.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. This work is supported in part by the US National Science Foundation under grants IIS-1924928, IIS-1938167 and OAC-1934600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[2] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

[3] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. 2003. On visual similarity based 3D model retrieval. In *Computer graphics forum*, Vol. 22. Wiley Online Library, 223–232.

[4] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[5] Zhengming Ding, Handong Zhao, and Yun Fu. 2018. *Learning representation for multi-view data analysis: models and applications*. Springer.

[6] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3558–3565.

[7] Yifan Feng, Zizhao Zhang, Xibin Zhao, Rongrong Ji, and Yue Gao. 2018. GVCNN: Group-view convolutional neural networks for 3D shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 264–272.

[8] iMotions. [n.d.]. *Facial expression analysis*.

[9] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[10] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[11] Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697

[12] Yan Li, Tao Yang, Jiayu Zhou, and Jieping Ye. 2018. Multi-Task Learning based Survival Analysis for Predicting Alzheimer's Disease Progression with Multi-Source Block-wise Missing Data. In *SDM*. 288–296.

[13] Mingxia Liu, Yue Gao, Pew-Thian Yap, and Dinggang Shen. 2017. Multi-Hypergraph Learning for Incomplete Multimodality Data. *IEEE journal of biomedical and health informatics* 22, 4 (2017), 1197–1208.

[14] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2247–2256.

[15] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[16] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 973–982.

[17] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 439–448.

[18] Chao Shang, Aaron Palmer, Jiangwen Sun, Ko-Shin Chen, Jin Lu, and Jinbo Bi. 2017. VIGAN: Missing view imputation with generative adversarial networks. In *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE.

[19] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*. 945–953.

[20] Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang. 2019. Metric learning on healthcare data with incomplete modalities. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 3534–3540.

[21] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1405–1414.

[22] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the 7th International Conference on Learning Representations*.

[25] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 949–954.

[26] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. 2018. Partial Multi-view Clustering via Consistent GAN. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE.

[27] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous Graph Attention Network. In *The World Wide Web Conference*. ACM, 2022–2032.

[28] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1912–1920.

[29] Shuo Xiang, Lei Yuan, Wei Fan, Yalin Wang, Paul M Thompson, and Jieping Ye. 2013. Multi-source learning with block-wise missing data for Alzheimer's disease prediction. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 185–193.

[30] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage* 61, 3 (2012), 622–632.

[31] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.

[32] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[33] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*. https://github.com/A2Zadeh/CMU-MultimodalSDK

[34] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[35] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous Graph Neural Network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 793–803.

[36] Lei Zhang, Yao Zhao, Zhenfeng Zhu, Dinggang Shen, and Shuiwang Ji. 2018. Multi-view missing data completion. *IEEE Transactions on Knowledge and Data Engineering* 30, 7 (2018), 1296–1309.

[37] Ziyuan Zhao, Huiying Zhu, Zehao Xue, Zhao Liu, Jing Tian, Matthew Chin Heng Chua, and Maofu Liu. 2019. An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management* 56, 6 (2019), 102097.

[38] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. 2007. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in neural information processing systems*. 1601–1608.

A APPENDIX

A.1 Data Descriptions

We conducted experiments on three datasets across two application domains. Datasets for the same application domain are collected from same sources.

3D Object Recognition. ModelNet40 [28] and NTU [3] datasets are used for this application domain. Following [6], the two input modalities are the two views of shape representations extracted from Multi-view Convolutional Neural Network (MVCNN) [19] and Group-View Convolutional Neural Network (GVCNN) [7]. Both the MVCNN and the GVCNN features are calculated by employing 12 virtual cameras to capture views with a interval angle of 30 degree.

Multimodal Emotion Recognition. For the IEMOCAP dataset, following [14, 31], we adopted the same feature extraction scheme for language, visual and acoustic modalities. *Language* features are obtained from the pre-trained 300-dimensional Glove word embeddings [15], which encode a sequence of transcribed words into a sequence of word vectors; *Visual* features are extracted as indicators of facial muscle movement, using Facet [8], include 20 facial action units, 68 facial landmarks, head pose, gaze tracking and HOG features; and, *Acoustic* features are obtained from time-series audio using the COVAREP acoustic analysis framework [4], including 12 Mel frequency cepstral coefficients (MFCCs), pitch, voiced/unvoiced segmentation, glottal source, peak slope, and so on. We obtain the above features from CMU Multimodal SDK [33], which can be accessed from //github.com/A2Zadeh/CMU-MultimodalSDK. Samples in the SDK are annotated according to the presence of four emotion categories (i.e., happy, sad, neutral and angry). For each emotion, we can conduct a binary classification task. In this paper, we conducted experiments on three of them (happy, sad and angry), as shown in Table 3. In other words, we trained a total of three models on this dataset.

A.2 Baseline Model Descriptions

We compared our method with the following six baselines.

Concat baseline performs fusion by concatenating unimodal features before a fully connected classifier. Our model use the same unimodal embedding networks as this baseline.

Tensor Fusion Network (TFN) [31] introduces tensor product mechanisms to model unimodal interactions. It is an inductive multimodal fusion model which is proposed without considering the existence of unexpected missing modalities.

Low-rank Fusion Network (LFM) [14] tries to approximate the expensive tensor products by performing efficient multimodal fusion with modality-specific low-rank factors, which is also an inductive learning method not designed for handling incompleteness.

Hypergraph Neural Network (HGNN) [6] studies deep graph learning and node classification in traditional hypergraph structures. It also applies the approach to multimodal prediction tasks, but simply concatenates unimodal features as the input node features. This method cannot deal with the heterogeneous and highly-interacted incomplete multimodal data.

Multi-task multimodal learning (MTL) baseline combines the proposed intra-hypernode encoder and pattern-specific classifiers. We directly use the original incomplete data to train this baseline without imputation. All multimodal instances share the same intra-hypernode encoder, but different patterns of instances use different classifiers. This baseline aims to test the impact of graph fusion strategy.

A.3 Baseline Reproducibility

The above baselines have public source code, but still require extra effort to fit our problem settings to their models. More details to implement baselines are as follows.

First, for the TFN, LFM, Concat and MTL baselines, the model sizes of unimodal networks and final-layer classifiers are the same as those in the proposed model. Other hyperparameters follow their original settings. Second, for the baselines that cannot deal with missing modaities (i.e., Concat, TFN, LFM, HGNN), we preprocess the input data by imputing the missing modalities with zero or values. We have also tested average imputation, but the performances of the baselines using average imputation are worse than those using zero imputation. Thus, we use zero imputation to perform all baselines in this paper. Third, for HGNN baseline, we preprocess the input data by concatenating all modalities in a node to shape proper feature vectors as input feature matrix. Note that for multimodal dataset may comes with 2D- or 3D-tensor sequential data/features (e.g., image, video, and audio features), we cannot concatenate them using the same way as 1D-tensor data. In order to apply HGNN on such tasks, we take the sum value over the additional dimensions, and then modalities can be concatenated. Graph edges were constructed using the same way in our work. Similar to the proposed model, we also stack two HGNN layers in all experiments.

Table 4: Hyperparameters for HHG graphs and MGMF models with bimodal and trimodal settings.

Hyper-parameters	HGMF (M=2)	HGMF (M=3)
k	10	10
L	2	2
$\mathcal{M}$	$\{1, 2\}$	$\{1, 2, 3\}$
$\mathcal{T}$	$\{1, 2, 3\}$	$\{1, 2, ..., 7\}$
$F_m, \forall m \in \mathcal{M}$	$\{128, 128\}$	$\{128, 128, 128\}$
F'	128	128
$d_p^0, \forall p \in \mathcal{T}$	$\{512 \vert \forall p \in \mathcal{T}\}$	$\{512 \vert \forall p \in \mathcal{T}\}$
$d_p^1, \forall p \in \mathcal{T}$	$\{128 \vert \forall p \in \mathcal{T}\}$	$\{128 \vert \forall p \in \mathcal{T}\}$
d	64	64
learning rate	1e-4	1e-3

A.4 Model Settings

We employed Pytorch3 to implement HGMF and all baselines, and conducted experiments on a single-core GPU. During graph construction, the hypernodes are associated with the original pre-extracted features. Each element in a hypernode can be of different dimension and are not concatenated; the language modality is in 3D-tensor format and others are 2D-tensor. The k for constructing the high-order hyperedges (in Section 3.1) equals 10 in each experiment. Note that as we let edge weights be 1, we construct a graph that only reflect data connection information.

There are many hyperparameters in the proposed model. The intra-hypernode encoder parameter set in Algorithm 1 can be summarised as

$$\theta_e = \{\mathbf{\Theta}_m, \mathbf{U}_{\mathcal{S}}, \mathbf{b}_{\mathcal{S}} | \forall m \in \mathcal{M}, \forall \mathcal{S} \in \mathcal{P}(\mathcal{M}) \setminus \emptyset\}, \tag{10}$$

Similarly, the MBGAT's parameter set can be represented as

$$\theta_g = \{\vec{\mathbf{a}}_p^{(l)}, \vec{\mathbf{b}}_p^{(l)}, \mathbf{W}_{pq}^{(l)}, \mathbf{V}_{pq}^{(l)} | \forall p, q \in \mathcal{T}, 0 \leq l < L\}. \tag{11}$$

We built up HGMF models in bimodal and trimodal settings. Table 4 summarises the hyperparameters of the HGMF models used in our experiments, including both graph structure and neural network hyperparameters. For intra-hypernode encoder, the embedding dimension of unimodal hidden representations are 64 for visual and language modalities, and 128 for other modalities, which are similar to those in baseline models. Encoded feature dimensions between different patterns can be significantly different. We let the dimension of each factor (an extracted modality-specific or interaction information) equal to 128. We stack two MBGAT layers as the transductive fusion stage of HGMF. At the first layer, we let each pattern's new feature space dimension is half of input dimension. the final embedding dimension for all patterns equal to 64, meaning that they are encoded into the same space.