iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله
مورد تایید ایران پیپر

Pattern Recognition 172 (2026) 112737

Contents lists available at ScienceDirect

Pattern Recognition

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

MKGPL: graph prompt learning with multi-view knowledge for few-shot recognition

Yanzhao Xie, Man Qiu, Yangtao Wang*, Siyuan Chen, Meie Fang, Maobin Tang, Wensheng Zhang

School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China

ARTICLE INFO

Keywords:
Large-scale models
Graph prompt
Few-shot recognition
Multi-view knowledge

ABSTRACT

Despite the strong generalization capability of large-scale models (LMs), their inherent massive parameters hinder rapid deployment and application in downstream task-specific scenarios. Consequently, parameter efficient fine-tuning (PEFT) methods have emerged as critical solutions. However, most existing approaches exhibit two primary limitations: (1) their inability to simultaneously integrate task-specific knowledge embedded in both positive and negative textual descriptions; and (2) the lack of exploration regarding the appropriate integration of graphs derived from downstream task data with graph prompts, resulting in suboptimal performance. To address these challenges, we propose an efficient Graph Prompt Leaning with Multi-view Knowledge (*i.e.*, MKGPL) framework that integrates a positive text subgraph, a negative text subgraph, and an image-specific subgraph for joint representation learning. This framework first constructs distinct subgraphs from features of corresponding images, positive and negative textual descriptions, enabling effective modeling of cross-modal knowledge while capturing information from both positive and negative text graph prompts. Specifically, the knowledge graph structure of our framework comprises three subgraphs, *i.e.*, image-specific subgraph, positive/negative text subgraph. The nodes correspond to image features, positive textual semantics, and negative textual semantics, respectively. Edges between nodes are constructed using widely adopted graph construction methods to represent and emphasize their correlations. Furthermore, we innovatively design graph-specific prompt tokens to enhance the framework's perceptual and adaptive capacity for downstream task data. These designs allow each image feature to perceive multi-view and multi-modal semantic information from diverse subgraphs, thereby generating more discriminative and high-performing classifiers for downstream tasks. Extensive experiments on 11 mainstream benchmark datasets for PEFT methods demonstrate that our MKGPL significantly outperforms previous state-of-the-art methods in various few-shot recognition tasks. Code is publicly available at https://github.com/MultimodalGra/MKGPL.

1. Introduction

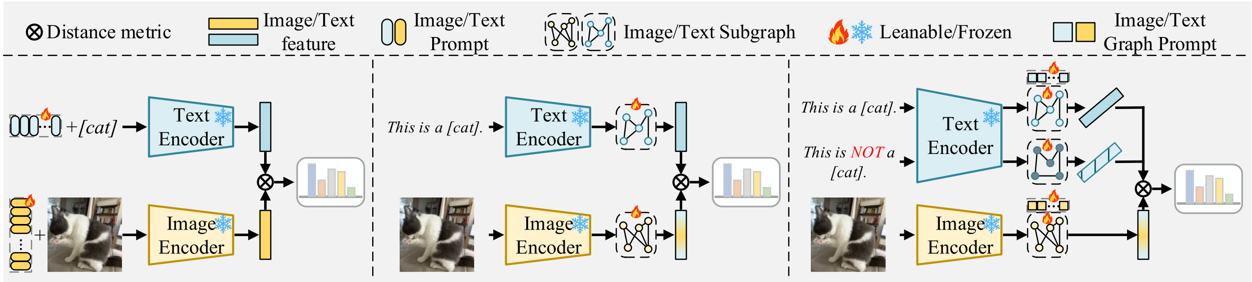
In recent years, the field of artificial intelligence has witnessed a remarkable proliferation of large-scale models (LMs) [1,2], whose immense knowledge reservoirs and superior generalization capabilities across diverse domains have garnered substantial research interest. To address the computational challenges associated with full-parameter fine-tuning on downstream tasks, emerging parameter-efficient fine-tuning (PEFT) [3–7] methodologies have been proposed to enable rapid adaptation of these pre-trained LMs through selective parameter modulation while preserving their foundational knowledge acquisition capabilities.

Existing PEFT paradigms can be primarily categorized into two families: 1) prompt-based contextual learning methods [3,4,7,8] and 2) adapter-based learning approaches [5,6,9,10]. The former typically involves injecting a sparse set of learnable vectors into downstream task data as contextual priors, where gradient updates optimize these vectors to align with task-specific distributions. The latter introduces lightweight parameter-efficient modules as pluggable components into pre-trained LMs, enabling fine-tuning to establish task-adaptive mappings between inputs and outputs. Both strategies achieve effective downstream adaptation of LMs while restricting parameter updates to minimal subsets. Both categories of methods effectively minimize training overhead for downstream adaptation of pre-trained models. This

[*] Corresponding author.
E-mail addresses: yzhx@gzhu.edu.cn (Y. Xie), qm123@e.gzhu.edu.cn (M. Qiu), ytaowang@gzhu.edu.cn (Y. Wang), chensiyuan@gzhu.edu.cn (S. Chen), fme@gzhu.edu.cn (M. Fang), tmb178@gzhu.edu.cn (M. Tang), wensheng.zhang@gzhu.edu.cn (W. Zhang).

https://doi.org/10.1016/j.patcog.2025.112737
Received 31 July 2025; Received in revised form 12 October 2025; Accepted 11 November 2025
Available online 16 November 2025
0031-3203/© 2025 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله
مورد تایید ایران پیپر



(a) Prompt Learning (b) GraphAdapter (c) Our method

Fig. 1. The comparison between (a) Classic Prompt Learning, (b) GraphAdapter, and (c) our proposed MKGPL. (a) Prompt learning methods directly integrate a limited number of learnable vectors (*e.g.*, via summation or concatenation) into input data, failing to exploit latent inter-data relationships. (b) GraphAdapter employs an adapter-based architecture, incorporating lightweight learnable graphs into pre-trained models while leveraging graph structures to utilize structural knowledge from positive text descriptions. (c) Our proposed MKGPL extends beyond prior methods by: (1) mining latent relationships between data instances; (2) incorporating both positive and negative textual knowledge; and (3) introducing adaptive learnable contextual components to graph structures, enabling dynamic knowledge integration.

efficiency arises because the trainable components are confined to either task-specific prompt contexts or lightweight learnable modules, with their parameter scales being orders of magnitude smaller than the parameters of the full pre-trained model. Such a design enables efficient fine-tuning while preserving the majority of pre-existing knowledge.

Among existing PEFT methodologies, several approaches have made some preliminary explorations in leveraging structural knowledge (*e.g.*, GraphAdapter [10]). These methods primarily focus on mining latent relationships between image samples and positive textual descriptions for subsequent utilization. Specifically, such approaches construct graph structures around image instances and their corresponding textual annotations, where nodes represent either visual features or positive textual descriptors, and edges encode internode associations typically quantified through distance metrics. By employing graph convolution operations, these methods enable information propagation across nodes within both image and text subgraphs, with edge weights governing the flow intensity of node-level features. This process facilitates implicit knowledge transfer through relational modeling, thereby exploiting cross-modal associations to guide model training and achieve effective few-shot classification performance.

However, existing methods combined with structure knowledge still face challenges in two aspects. Firstly, they lack the ability to simultaneously explore the internal structural knowledge of both positive and negative textual descriptions. This leads to a situation where cross-modal matching only focuses on iteratively narrowing the distance between image features and positive text features, while neglecting the necessity of enlarging the distance between image features and negative text features. Secondly, although the integration of graph structures can enhance the few-shot classification performance and downstream task generalization capability of vision-language pre-trained models, the absence of corresponding prompt information within the graph structure itself would result in suboptimal overall model performance. Building upon the aforementioned analysis, we conduct a methodological comparison among classic prompt learning paradigms, graph-based adapter approaches, and our proposed MKGPL, with visualizations presented in Fig. 1.

To address the first challenge, we propose a novel graph prompt based training approach, termed MKGPL. This method is designed to model task-specific structural knowledge for downstream task data by jointly considering image features, positive textual features, and negative textual features. Beyond merely leveraging these individual feature representations, our approach emphasizes mining the interrelationships among them to enable selective information aggregation. Specifically, we tackle two critical technical challenges by graph structure: 1) developing mechanisms to capture structured knowledge representations, and 2) exploring effective inter-modal information fusion strategies across the modalities of images and positive/negative texts.

To deal with the second challenge, we introduce universal graph prompt [11,12] to overcome catastrophic forgetting [13] or overfitting issues when downstream task data is scarce, we adhere to the prompt learning paradigm by introducing a small number of learnable parameters into the graph structure. Additionally, designing feasible prompting functions remains highly challenging in the graph pre-training domain due to the absence of standardized pre-training tasks. While pioneering works [12,14] have attempted to apply prompt-based tuning methods to edge prediction [11] pre-trained models by incorporating virtual class-prototype nodes/graphs with learnable connections into original graphs (*i.e.*, making the adaptation process resemble edge prediction), these approaches exhibit limited applicability and are only compatible with specific model architectures. When confronted with more complex pre-training strategies, manually designing prompting functions following the link prediction paradigm becomes extremely difficult, preventing direct migration of existing prompt-based tuning methods. Furthermore, by integrating the learnable vectors into the extracted feature structures following the TaskRes [5], we introduce a universally applicable prompt learning strategy for graph structures in our proposed approach. These graph prompts operate on the input graph's feature space, implementing shared learnable vectors appended to all node features within the graph.

Our contributions can be summarized as below:

- We propose a novel graph-structured prompt learning approach named MKGPL to address few-shot learning challenges for large-scale models in downstream specialized tasks. Our method constructs distinct subgraphs for image data, and positive/negative textual descriptions respectively, enabling structured feature mining across heterogeneous data modalities.
- To mitigate catastrophic forgetting and overfitting issues in pre-trained models under few-shot scenarios, we introduce a universal graph prompting strategy. Specifically, lightweight learnable prompt features are incorporated into both image-specific and positive/negative text subgraphs, enhancing the model's few-shot generalization capability through cross-modal and inter-modal information fusion.
- Extensive experiments on 11 benchmark datasets demonstrate that our MKGPL consistently outperforms existing state-of-the-art (SOTA) methods across various evaluation metrics. For instance, MKGPL achieves a 1.97 % improvement over the existing SOTA approaches on Food101 dataset under the 1-shot setting.

2. Related works

2.1. Visual-language models

Visual-language models (VLMs) are designed to leverage aligned image-text pair information through architectures such as Trans-

iranpaper Downloaded from https://iranpaper.ir

https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

former [2] or ResNet [2,15], constructing task-specific text/image encoders and formulating tailored loss functions. This paradigm aims to capture the inherent generalizable semantic information embedded in large-scale multimodal datasets. Recent studies have increasingly leveraged the generalizable multimodal features extracted by VLMs to enhance task-specific performance across diverse domains, such as few-shot learning [5,9,10], cross-modal generation [16,17], and image recognition [15,18]. The foundational work in this paradigm is established by CLIP [2], which pioneered the use of contrastive loss functions and leveraged conventional visual-textual encoders to extract features and achieve cross-modal semantic alignment. VLMs can be systematically categorized into two primary paradigms based on their architectural focus, *i.e.*, 1) dual-branch encoder architectures [2] and 2) cross-modal fusion encoder frameworks [19]. The former employs independent encoding pathways for visual and textual modalities, maintaining modality-specific processing before late alignment/fusion, while the latter integrates multimodal interactions through early fusion mechanisms or cross-attention mechanisms that enable joint representation learning. This categorization reflects fundamental design choices in balancing modality-specific feature extraction with cross-modal semantic fusion, directly impacting downstream task performance in multimodal understanding and generation scenarios.

2.2. Graph learning

Graph Neural Networks (GNNs) [20–23] have garnered significant attention in the field of graph learning research, primarily due to their exceptional capability to model node relationships, capture topological structures, and generalize effectively for structured knowledge extraction across diverse graph-based tasks. GNNs are designed to model and learn structured knowledge through the lens of node features, edge connectivity patterns, and local-global relational dynamics. Early seminal works [24] such as Graph Convolutional Networks (GCNs) [20] emulate convolutional operations from CNNs, iteratively stacking convolutional layers to aggregate neighboring node features based on edge connectivity patterns. Graph Isomorphism Network (GIN) [25] introduces a theoretical framework to analyze the expressivity of Graph Neural Networks (GNNs) across diverse graph structural scenarios.

Recent years have witnessed a growing trend of integrating GNNs into VLMs learning to enhance structured feature extraction capabilities [26,27]. However, existing approaches either introduce substantial computational overhead through multi-perspective modeling frameworks (*e.g.*, model parameter perspective [26]), or employ hypergraph-based frameworks that unify image and text features into homogeneous graph structures, where complex topological configurations may impair model adaptability in few-shot learning scenarios. The present study proposes a novel approach leveraging simple unimodal graph architectures combined with GCN to systematically extract intra-modal structural features. Furthermore, we construct positive/negative text subgraphs and an image-specific subgraph to facilitate cross-modal information fusion while maintaining computational efficiency and structural interpretability. This design effectively balances the need for structured knowledge representation with the practical constraints of multimodal adaptation in resource-limited scenarios.

2.3. Parameters efficient fine-tuning

Parameters efficient fine-tuning (PEFT) methods focus on investigating strategies to efficiently fine-tune critical parameters in large-scale pre-trained models, aiming to enhance their adaptability and performance on downstream tasks. These critical parameters can be manifested as either appended learnable prompts [3,4,7,28] integrated into the input downstream data, or alternatively as a limited number of parameters contained within newly introduced lightweight learnable modules or components (*i.e.*, adapter) [5,9,10] added to the base model. Existing PEFT methods can be broadly classified into two primary categories, *i.e.*, prompt-based approaches and adapter-based methods. Prompt-based approaches, originating from Natural Language Processing (NLP), have been widely used to improve the adaptation of pre-trained large-scale models to diverse downstream tasks [19,29,30]. In recent years, an increasing number of prompt learning approaches have been applied to vision-text pre-trained models. As a pioneering effort, CoOp [3] and CoCoOp [4] respectively proposed replacing manually designed prompt templates (*i.e.*, “A photo of a” used in Clip-Adapter [9]) with learnable prompt vectors, with the latter further introducing a simple yet efficient component to map image features extracted from the visual branch onto these learnable vectors. MaPLe [31] incorporates an additional mapping network that utilizes a unidirectional data flow mechanism to strengthen cross-modal alignment between visual and textual prompt representations. This design significantly enhances the VLM’s capacity for downstream task adaptation through improved inter-modality information coupling. In addition to prompt learning approaches for VLMs, an increasing number of graph-structured prompt learning methods and empirical researches [12,14,32–34] have been proposed. Sun et al. [32] unified the formats of graph prompts and language prompts through prompt tokens, token structures, and insertion patterns, thereby realizing a novel multi-task prompting method for graph models. Recently, Wang et al. [33] provide an in-depth analysis of the underlying principles of graph prompting methods and points out the promising and reliable application prospects. Both of the aforementioned works mention a pivotal foundational work, GPF [12], which employs a simple yet computationally efficient strategy to directly incorporate prompt information into graph nodes through summation. Sun et al. [34] clearly pointed out that GPF can effectively handle graph classification tasks and align the objectives of task head fine-tuning and prompt fine-tuning. This approach enhances the few-shot learning capabilities of graph-based methods while maintaining low computational complexity.

Unlike conventional prompt learning approaches centered around VLMs, the adapter-based methodology aims to adapt pre-trained image/text features to align VLMs with downstream specific tasks. Representative works such as TaskRes [5] explicitly decouple the preservation of prior knowledge in pre-trained VLMs from task-specific knowledge acquisition without excessive reliance on pre-trained features. However, this approach falls short in effectively integrating structured knowledge. Subsequent GraphAdapter [10] incorporates structural knowledge but lacks adequate attention to downstream task data adaptation. Building upon these insights, this paper proposes a novel framework that synergizes positive/negative pre-trained textual features with image pre-training information, constructing graph-structured representations while introducing computationally efficient graph-based prompts. This design enables seamless adaptation to downstream tasks through structured knowledge infusion and task-aware feature recalibration, addressing the limitations of prior methods in simultaneous structural knowledge utilization and task-specific optimization.

3. Proposed method

In this section, we present our proposed method that comprehensively models structural knowledge across heterogeneous modalities through the integration of a triplet knowledge graph. Comprising a positive text subgraph, a negative text subgraph, and an image-specific subgraph, this tripartite architecture enables structured representation learning from complementary information sources.

3.1. Overview

The proposed MKGPL consists of three primary components: (1) a multi-encoder architecture for structured knowledge extraction across heterogeneous modalities, (2) a graph convolutional network (GCN) module that performs knowledge refinement through cross-subgraph

iranpaper Downloaded from https://iranpaper.ir https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر ترجمانو

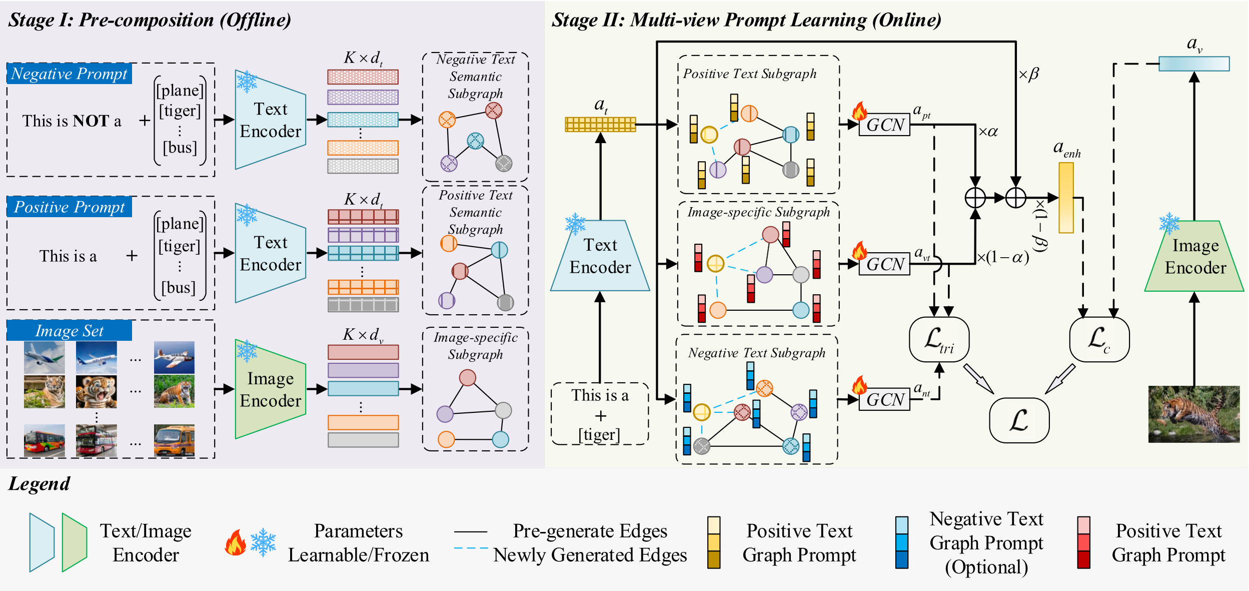


Fig. 2. An overview of our proposed MKGPL. The proposed framework operates through two sequential phases. 1) Pre-composition Phase (Offline): In this phase, an image/text encoder is utilized to extract features from a small number (*e.g.*, 1-/2-/4-/8-/16-shot) of manually labeled samples. Subsequently, static positive/negative text subgraphs and an image-specific subgraph are constructed. These pre-built subgraphs will remain fixed after initialization. 2) Multi-view Prompt Learning Phase (Online): This phase aims to integrate positive/negative text features and image features into a unified textual representation through multi-modal fusion. A small number of labeled samples are still required to participate in this training process. Additionally, lightweight graph prompts are introduced to enhance the model's adaptability and generalization ability while maintaining extremely low computational overhead. Finally, the inference process is similar to the stage II, except that the input image data is unlabeled, while the input text data consists of sentences formed by combining all possible labels with human-crafted prompts.

propagation, followed by loss computation via contrastive learning objectives. and (3) the universal graph prompt learning mechanism, which facilitates adaptive knowledge fusion through parameter-efficient tuning of graph-structured prompts while preserving domain-specific structural patterns. The whole framework of our MKGPL is depicted in Fig. 2.

The proposed MKGPL can be roughly divided into two stages, namely 1) the pre-composition stage and 2) the multi-view prompt learning stage. Specifically, the first stage primarily focuses on how to reasonably construct positive text subgraph, negative text subgraph, and image-specific subgraph. For few-shot learning, the amount of labeled data available to the model is relatively limited. Therefore, it is necessary to effectively exploit the internal correlations among these labeled data by leveraging the graph structure. In the second stage, after the graph construction is completed, various possible general graph prompts are injected into the graph structure. These lightweight prompt vectors are then automatically updated through backpropagation to enhance the model's overall domain adaptability and generalization performance. In the subsequent sections, we elaborate on the two-stage framework of the proposed MKGPL in detail.

3.2. Pre-composition

The first stage of our proposed method is the pre-composition. In this phase, we adhere to the common experimental setup of few-shot learning. The focus is on a limited set of labeled samples (*i.e.*, 1/2/4/8/16-shot) within the training dataset, which includes image samples along with their corresponding textual descriptions. Specifically, for the textual descriptions of each image, we design its negative counterparts. For instance, given an image containing a cat, the positive textual description could be "This is a [cat]", while the corresponding negative textual description would be "This is **NOT** a [cat]". During this phase, we construct separate subgraph structures centered around the image, as well as the positive and negative textual descriptions, respectively.

Text subgraph. Let $\mathbf{G} = \{\mathbf{G}_{nt}, \mathbf{G}_{pt}, \mathbf{G}_{v}\}$ denote the set composed of three knowledge subgraphs: the negative text subgraph $\mathbf{G}_{nt}$, the positive text subgraph $\mathbf{G}_{pt}$, and the image-specific subgraph $\mathbf{G}_{v}$. These three subgraphs respectively encode textual structural knowledge and image structural knowledge, forming a heterogeneous knowledge representation system that integrates cross-modal structural information. Subsequently, the image/text features can adaptively refine the structural knowledge of downstream tasks across dual modalities to achieve self-optimization, thereby leveraging tribrach structural knowledge to construct an enhanced model for downstream tasks.

To model structured knowledge representations of positive/negative descriptive texts, specifically the relationships between semantic nodes with heterogeneous meanings, we construct two specialized knowledge subgraphs: the positive text subgraph $\mathbf{G}_{pt} = \{\mathcal{N}_{pt}, \mathbf{E}_{pt}\}$ and the negative text subgraph $\mathbf{G}_{nt} = \{\mathcal{N}_{nt}, \mathbf{E}_{nt}\}$. Leveraging the inherent characteristics of the Contrastive Language-Image Pre-training (CLIP [2]) framework, where its text encoder generates class-discriminative features through contrastive learning across different category prompts, we observe that textual features within the same class prompt group effectively capture the semantic attributes of that category. Taking the positive text subgraph as an exemplar, node features are acquired by inputting artificial template prompts (*e.g.*, "This is a [classname]") into the pre-trained CLIP text encoder, which projects these inputs into a multimodal embedding space. For a classification task with K categories, both $\mathbf{G}_{pt}$ and $\mathbf{G}_{nt}$ contain K nodes corresponding to each category. Their respective node feature sets are formally represented as $\mathcal{N}_{pt} = \{n^{i}_{pt}\}^{K}_{i=1} \in \mathbb{R}^{K\times D}$ and $\mathcal{N}_{nt} = \{n^{i}_{nt}\}^{K}_{i=1} \in \mathbb{R}^{K\times D}$. Each D-dimensional feature vector encapsulates semantic information extracted via CLIP's cross-modal alignment mechanism, enabling structured cross-modal knowledge transfer. Notably, due to the similarity in construction methodologies between positive and negative text subgraph, the following discussion will primarily focus on the edge computation and graph construction process of the positive text subgraph.

After obtaining the node features, it becomes imperative to select an appropriate distance metric for quantifying the correlation strength between heterogeneous nodes. Considering that the CLIP inherently utilizes cosine similarity as the standard measurement for assessing dis-

iranpaper Downloaded from https://iranpaper.ir https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر ترجمانو

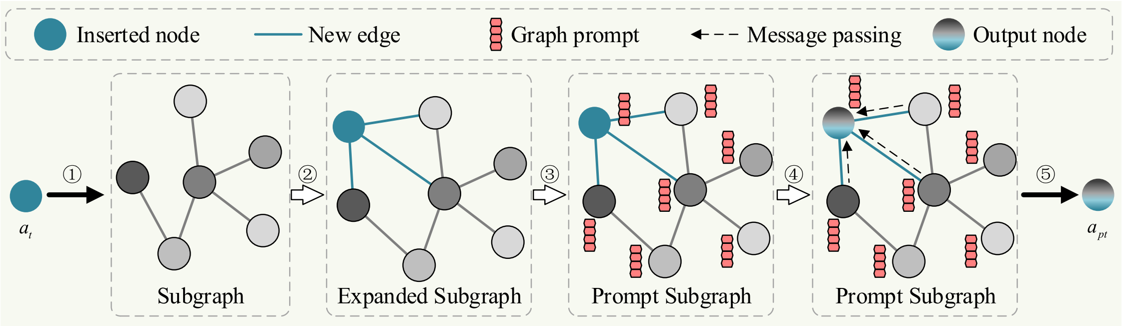


Fig. 3. An overview of inserting text nodes into a subgraph and performing information fusion (illustrated with the example of the positive text subgraph). The workflow proceeds as follows: ① A text feature is injected into a pre-composition subgraph structure; ② Edges are dynamically established between nodes based on pairwise similarity metrics; ③ Learnable graph prompts are inserted to enhance relational representations; ④ Graph convolution operations are applied via a GNN to propagate and fuse contextual information across the whole graph; ⑤ Matrix slicing is employed to extract the refined text feature as output after information fusion. This pipeline demonstrates the integration of structural graph learning with feature refinement in few-shot scenarios.

tances between cross-modal feature vectors in its contrastive learning paradigm, the proposed MKGPL also adopts the cosine distance metric during the construction of positive text subgraph. Let square matrix $\mathbf{E}_{pt} = \{e^{i,j}_{pt}\}$ denote the edge set, where each element is formally defined through the following computational formulation:

$$e^{i,j}_{pt} = cos(n^i_{pt}, n^j_{pt}) = \frac{n^i_{pt} \cdot n^{j\top}_{pt}}{||n^i_{pt}||_1 \cdot ||n^j_{pt}||_1}, i, j \in [1, K], \tag{1}$$

where $e^{i,j}_{pt}$ denotes the cosine distance between the features of i-th node (*i.e.*, n^i_{pt}) and j-th node (*i.e.*, n^j_{pt}), representing the edge connectivity between these two nodes. $|| \cdot ||_1$ means 1-norm. The negative text subgraph construction follows the same process.

Image-specific subgraph. The construction of the image-specific subgraph follows a two-component framework comprising node feature encoding and edge computation, denoted as $\mathbf{G}_v = \{\mathcal{N}_v, \mathbf{E}_v\}$. Specifically, node features $\mathcal{N}_v = \{n^i_v\}^K_{i=1} \in \mathbb{R}^{K\times D}$ are extracted by encoding images from the downstream dataset using CLIP's image encoder. Notably, this feature acquisition process differs from that of textual nodes. As detailed in Fig. 2, we first input images of the same class along with their augmented (*i.e.*, flipping, random crop, etc.) counterparts into the encoder to extract features for each image instance within the category. Subsequently, the mean of these features is computed to form the node set $\mathcal{N}_v$ of the image-specific subgraph. Similar to Eq. (1), the edge set $\mathbf{E}_v = \{e^{i,j}_v | i, j \in [1, K]\}$ is constructed by calculating the cosine distance between visual node features.

3.3. Multi-view prompt learning

Following the standard few-shot learning paradigm (*i.e.*, 1/2/4/8/16-shot settings), our approach maintains a task-specific knowledge subgraph construction process that is pre-constructed in a one-time manner and remains immutable for a given downstream dataset. Upon establishing the three knowledge subgraphs $\mathbf{G}_{pt}$, $\mathbf{G}_{nt}$, and $\mathbf{G}_v$ as described, we introduce a dedicated adaptation mechanism designed to enable multi-view collaborative learning between textual features and the image-specific subgraphs. This adaptation process systematically promotes cross-modal interaction through structured feature alignment and progressive knowledge aggregation, ensuring complementary information fusion across heterogeneous semantic spaces while preserving domain-specific knowledge boundaries.

In contrast to previous works that typically model either task-specific knowledge or single-modality data in isolation, resulting in suboptimal cross-modal information fusion, our approach simultaneously addresses intra-modality data fusion within textual modalities and inter-modality fusion between textual and visual modalities. Specifically, given a textual utterance encoded as feature vector a_t through text encoder of pretrained CLIP, our method MKGPL strategically injects this textual feature into three distinct subgraphs, *i.e.*, positive text subgraph, negative text subgraph, and image-specific subgraph. The fusion of text features with information from three subgraphs (positive text, negative text, and visual subgraphs) is abstracted into the workflow depicted in Fig. 3. This dual-purpose injection serves two objectives, including 1) preparing for cross-modal interactions between textual data with both positive/negative textual subgraph and visual subgraph, and 2) enabling effective aggregation and extraction of relational components from semantically similar nodes through graph topology, thereby enhancing the target textual features with structured contextual information. Taking the positive text subgraph as an example, the aforementioned process can be formally expressed as follows:

$$\mathcal{N}_{eppt} = [a_t, \mathcal{N}_{pt}], \quad \mathbf{E}_{eppt} = \begin{bmatrix} 1 & \cos(a_t, \mathcal{N}_{pt}) \\ \cos(a_t, \mathcal{N}_{pt}) & \mathbf{E}_pt \end{bmatrix}, \tag{2}$$

where $\mathcal{N}_{eppt}$ denotes expanded node feature matrix, $\mathbf{E}_{eppt}$ represents the expanded adjacency matrix. Following analogous processing, we can derive the expanded node feature matrices and expanded adjacency matrices for both the negative text subgraph and the image-specific subgraph, *i.e.*, $\mathcal{N}_{epnt}$ and $\mathcal{N}_{epv}$, $\mathbf{E}_{epnt}$ and $\mathbf{E}_{epv}$.

Following the acquisition of augmented node feature matrices and adjacency matrices, we further introduce universal graph prompts to enhance the robustness of graph-augmented textual features. Specifically, we inject learnable universal graph prompts $p \in \mathbb{R}^D$ into each node within the positive text subgraph and image subgraph. Following the graph prompt methodology [12], we adopt a direct summation approach to integrate these trainable graph prompts with node representations. Taking the image subgraph as an example, this augmentation process is formalized as: $\mathbf{E}_{epv} = [a_t + p, n^1_v + p, n^2_v + p, \dots, n^K_v + p]$. Subsequently, we employ a Graph Convolutional Network (GCN) to propagate and update node representations, enabling fusion of a_t features with those of connected neighbor nodes. This propagation mechanism facilitates cross-modal knowledge transfer across three distinct modalities: positive textual descriptions, negative textual descriptions, and visual content. The complete feature integration process can be mathematically formulated as:

$$\begin{aligned} \tilde{\mathcal{N}}_{eppt} &= GCN_{pt}(\mathcal{N}_{eppt}, \hat{\mathbf{E}}_{eppt}) = \sigma(\hat{\mathbf{E}}_{eppt}\mathcal{N}_{eppt}W_{pt}), \\ \tilde{\mathcal{N}}_{epnt} &= GCN_{nt}(\mathcal{N}_{epnt}, \hat{\mathbf{E}}_{epnt}) = \sigma(\hat{\mathbf{E}}_{epnt}\mathcal{N}_{epnt}W_{nt}), \\ \tilde{\mathcal{N}}_{epv} &= GCN_{v}(\mathcal{N}_{epv}, \hat{\mathbf{E}}_{epv}) = \sigma(\hat{\mathbf{E}}_{epv}\mathcal{N}_{epv}W_{v}), \end{aligned} \tag{3}$$

where $GCN(\cdot)$ denotes the graph convolutional layer operation corresponding to the specific subgraph, $\sigma(\cdot)$ represents the activation function, W corresponds to the learnable weight matrix within each graph convolutional layer across the corresponding subgraph. In addition, $\hat{\mathbf{E}}_{eppt} = D_{eppt}\mathbf{E}_{pt}D_{eppt}$, $\hat{\mathbf{E}}_{epnt}D_{epnt}\mathbf{E}_{nt}D_{epnt}$ and $\hat{\mathbf{E}}_{epv} = D_{epv}\mathbf{E}_vD_{epv}$ respectively denote adjacency matrices specifically optimized for graph convolution operations, while $D_{eppt} = diag(\sum^K_{j=1}(\mathbf{E}_{pt} + I)_j)$, $D_{epnt} =$

Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

$diag(\sum_{j=1}^{K}(\mathbf{E}_{nt}+I)_j)$ and $D_{epv}=diag(\sum_{j=1}^{K}(\mathbf{E}_v+I)_j)$ respectively represent edge-wise Laplacian normalization [35] applied to the graph structure. Following the graph convolution operations, the newly incorporated textual feature node a_t integrates feature representations from neighboring nodes across corresponding subgraphs. Subsequently, this enhanced textual feature needs to be extracted through an indexing operation. Specifically, this is achieved by slicing the aggregated feature matrix, *i.e.*, $a_{pt}=\tilde{\mathcal{N}}_{eppt}[0,:]$, $a_{nt}=\tilde{\mathcal{N}}_{epnt}[0,:]$ and $a_{vt}=\tilde{\mathcal{N}}_{epvt}[0,:]$. The resulting features can be interpreted as augmented textual representations, encapsulating original textual features, positive/negative textual information, and visual information from the corresponding subgraph. This design facilitates both intra-modal feature refinement and cross-modal information fusion. To further unify these multi-view features, we introduce a learnable hyperparameter α that balances intra-modal and cross-modal contributions through weighted summation, *i.e.*, $a_t^*=\alpha a_{pt}+(1-\alpha)a_{vt}$. Finally, to prevent the original textual features from being overly diluted in the final textual representation, we explicitly propose an element-wise summation strategy that aggregates the aforementioned enhanced textual features a_t^* with the original textual features a_t, *i.e.*, $a_{enh}=\beta a_t^*+(1-\beta)a_t$. $\beta\in[0,1]$ serves as a learnable balancing factor to control the contribution ratio between enhanced and original features.

3.4. Objective function

The objective function of our proposed MKGPL comprises two primary components, *i.e.*, 1) a contrastive loss item $\mathcal{L}_c$ between image features and augmented text features, and 2) a triplet loss $\mathcal{L}_{tri}$ among text features fused with positive/negative text information and image information respectively. For the former item, we incorporate a contrastive loss analogous to that employed in the CLIP [2] pre-trained model. The mathematical formulation of the contrastive loss is presented as follows:

$$\mathcal{L}_c=-log\frac{exp(sim(a_v,a_{enh}^{+})/\tau)}{\sum_{j=1}^{K}exp(sim(a_v,a_{enh}^{k}))/\tau},\tag{4}$$

where a_{enh}^{+}, sim and τ respectively represent the enhancement text vector corresponding to query vector (*i.e.*, image feature a_v), cosine similarity and learned temperature of CLIP. The triplet loss $\mathcal{L}_{tri}$ is formulated around text features embedded with structural information from three subgraphs. Its core objective is to enforce proximity between image-fused text representations (denoted as a_{vt}) and positive text-conditioned counterparts (denoted as a_{pt}) in the embedding space, while simultaneously maximizing the distance from negative text-conditioned representations (denoted as a_{nt}). This mechanism effectively reduces feature distances for semantically aligned information pairs and increases separation for inversely correlated representations. The mathematical formulation is presented as follows:

$$\mathcal{L}_{tri}=\sum_{i=1}^{N}max(||a_{vt}^{i}-a_{pt}^{i}||_2^2-||a_{vt}^{i}-a_{nt}^{i}||_2^2+m,0),\tag{5}$$

where N denotes the number of samples, $||\cdot||_2$ denotes the L2 norm operation, and m denotes the margin that enforces a minimum separation of m between the L2-normalized distances of positive and negative feature pairs. The overall objective function of the proposed method MKGPL can thus be formulated by integrating the aforementioned components and a balance weight γ, expressed mathematically as $\mathcal{L}=\mathcal{L}_c+\gamma\mathcal{L}_{tri}$. During the inference phase, MKGPL preserves the three subgraph structures tailored to each specific dataset, along with the associated GCN model and its pre-trained parameters. Notably, the input image data in this stage are devoid of any labels. Instead, following an encoding process and subsequent information aggregation through the GCN, similarity computations are conducted between the encoded image features and the text features representing potential categories. Ultimately, the text description exhibiting the highest degree of similarity is designated as the predicted category for the input image.

4. Experiments

4.1. Experimental settings

Datasets. We conducted extensive experiments on few-shot classification and domain generalization tasks to evaluate the effectiveness of our proposed MKGPL. Following previous prompt-based studies [4,8,10], we selected 11 benchmark datasets for few-shot classification tasks, including ImageNet [36], StanfordCars [37], UCF101 [38], Caltech101 [39], Flowers102 [40], SUN397 [41], DTD [42], EuroSAT [43], FGVCAircraft [44], OxfordPets [45], and Food101 [46]. Among them, OxfordPets, Food101, StanfordCars, Flowers102, and FGVCAircraft belong to fine-grained classification datasets. EuroSAT is a remote sensing image classification dataset. DTD specializes in texture classification. For few-shot learning evaluations, we compared model performance across 1/2/4/8/16-shot settings while maintaining full test set evaluations. Following the experimental protocols established in GraphAdapter [10], we constructed generalization experiments using the ImageNet-V2 [47], ImageNet-Sketch [48], ImageNet-A [49], and ImageNet-R [50] datasets.

Implementation details. For a fair comparison, our experiments are conducted using the CLIP-pretrained ResNet-50 backbone by default, consistent with prior methods. We employ a universal text prompt template to generate positive textual descriptions for each class. Negative text prompts are constructed by inserting “NOT” to the original positive prompts. Our MKGPL is fine-tuned using the Adam optimizer with cosine learning rate decay. Notably, we adopt the warm-up strategy from previous works during training, starting from an initial learning rate of $1e-5$ to ensure stable training in the initial epoch. Visual subgraphs are constructed using few-shot training samples. Before the training, we leverage the pre-trained CLIP visual encoder to extract visual features from these few-shot samples within each class, which are then averaged to form class-specific node representations. The number of images per class corresponds directly to the shot number specification (e.g., four images per node for 4-shot training). The data augmentation pipeline exclusively includes “random resized cropping” and “random horizontal flipping” operations.

To enhance the reproducibility and robustness of experimental results, we conducted three independent trials with differentiated random seeds in each experimental iteration, and the mean value of outcomes across these trials was adopted as the final results. All training and inference processes were executed exclusively on a single NVIDIA RTX A40 GPU. Specifically, when performing large-scale graph structure training tasks on the ImageNet dataset, we employed a subgraph decomposition strategy that partitioned the original 1000-node graph into four independent subgraph modules (each containing 256 nodes). This approach maintained the structural integrity of the graph while effectively reducing the computational complexity of individual training iterations.

4.2. Performance comparison with SOTA methods

Few-shot learning. We compare our MKGPL with SOTA methods across 11 benchmark datasets, including CoOp [3], TaskRes [5], Tip-Adapter [6], CLIP-Adapter [9], and GraphAdapter [10]. As shown in Table 1, our method consistently outperforms previous works under 1-/2-/4-/8-/16-shot settings across all 11 benchmarks. Notably, in the 1-shot configuration, our approach achieves an average performance of 65.54 %, surpassing GraphAdapter by 0.74 % and TaskRes by 1.50 %. This validates the advantage of learning positive and negative textual knowledge in downstream tasks. Observing results across datasets with varying class cardinalities, including large-scale class distributions (*e.g.*, 1000 classes in ImageNet) and small-scale class configurations (*e.g.*, 10 classes in EuroSAT), our MKGPL demonstrates competitive performance, maintaining superiority over baseline methods across all shot configurations from 1-shot to 16-shot on both ImageNet and EuroSAT

iranpaper Downloaded from https://iranpaper.ir

https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر ترجمانو

Table 1
Comparisons of few-shot learning results with ResNet-50 backbone. (The gray shading and the **bold** values respectively represent MKGPL results and the best one. The underline values denote the suboptimal performance).

Methods	Setting	Caltech101	DTD	EuroSAT	FGVCAircraft	Flowers102	Food101	ImageNet	OxfordPets	StanfordCars	SUN397	UCF101	Avg.
Zero-shot CLIP	1-shot	86.27	42.32	37.56	17.28	66.14	77.31	58.18	85.77	55.61	58.52	61.46	58.77
CoOp		87.53	44.39	50.63	9.64	68.12	74.32	57.15	85.89	55.59	60.29	61.92	59.59
Clip-Adapter		88.60	45.80	61.40	17.49	73.49	76.82	61.20	85.99	55.13	61.30	62.20	62.67
Tip-Adapter-F		88.80	50.49	50.34	19.01	**81.17**	76.22	60.88	86.04	56.78	61.23	**66.19**	63.38
TaskRes		88.80	50.17	61.27	21.20	78.77	74.03	61.43	83.50	58.77	61.93	64.57	64.04
GraphAdapter		88.90	**51.77**	63.30	20.93	79.98	75.43	61.50	84.40	59.70	61.93	64.93	64.80
MKGPL (Ours)		**89.20**	51.17	**63.57**	**21.70**	79.75	**77.40**	**61.83**	**86.43**	**61.33**	**62.76**	65.69	**65.54**
Zero-shot CLIP	2-shot	86.29	42.32	37.56	17.28	66.14	77.31	58.18	85.77	55.61	58.52	61.46	58.77
CoOp		87.93	45.15	61.50	18.68	77.51	72.49	57.81	82.64	58.28	59.48	64.09	62.32
Clip-Adapter		89.37	51.48	63.90	20.10	81.61	77.22	61.52	86.73	58.74	63.29	67.12	65.55
Tip-Adapter-F		89.61	55.32	64.76	21.76	85.40	77.05	61.57	86.06	61.13	63.19	68.99	66.80
TaskRes		90.03	54.53	65.77	23.07	**85.63**	75.30	62.17	84.43	62.77	64.33	69.10	67.02
GraphAdapter		90.20	**55.75**	**67.27**	23.80	**85.63**	76.27	62.32	86.30	63.23	64.60	69.47	67.71
MKGPL (Ours)		**90.37**	55.18	**67.27**	**24.07**	85.29	**77.47**	**62.39**	**87.47**	**63.97**	**64.80**	**69.91**	**68.01**
Zero-shot CLIP	4-shot	86.29	42.32	37.56	17.28	66.14	77.31	58.18	85.77	55.61	58.52	61.46	58.77
CoOp		89.55	53.49	70.18	21.87	86.20	73.33	59.99	86.70	62.62	63.47	67.03	66.77
Clip-Adapter		89.98	56.86	73.38	22.59	87.17	77.92	61.84	87.46	62.45	65.96	69.05	68.61
Tip-Adapter-F		90.87	**60.25**	69.66	26.39	89.53	77.46	62.62	86.46	64.86	65.88	**72.71**	69.70
TaskRes		90.63	59.50	72.97	24.83	89.50	76.23	62.93	86.27	66.50	66.67	69.70	69.61
GraphAdapter		90.97	59.63	75.20	26.97	**89.90**	76.77	63.12	86.57	66.53	66.70	71.47	70.35
MKGPL (Ours)		**91.73**	59.77	**75.33**	**27.17**	89.33	**78.34**	**63.23**	**87.70**	**66.83**	**66.83**	71.70	**70.72**
Zero-shot CLIP	8-shot	86.29	42.32	37.56	17.28	66.14	77.31	58.18	85.77	55.61	58.52	61.46	58.77
CoOp		90.21	59.97	76.63	26.13	91.18	71.82	61.56	85.32	68.43	65.52	71.94	69.89
Clip-Adapter		91.40	61.00	77.93	26.25	91.72	78.04	62.68	87.65	67.89	67.50	73.30	71.40
Tip-Adapter-F		91.70	62.93	79.33	30.62	91.00	77.90	64.15	**88.28**	69.51	69.23	74.76	72.67
TaskRes		92.23	64.23	78.07	29.50	**94.30**	76.90	64.03	87.07	70.57	68.70	74.77	72.76
GraphAdapter		**92.45**	64.50	80.17	31.37	94.07	77.73	64.23	87.63	70.53	68.97	75.73	73.40
MKGPL (Ours)		92.35	**64.70**	**80.40**	**31.87**	94.17	**78.53**	**64.37**	87.93	**71.30**	**69.33**	**76.27**	**73.75**
Zero-shot CLIP	16-shot	86.29	42.32	37.56	17.28	66.14	77.31	58.18	85.77	55.61	58.52	61.46	58.77
CoOp		91.83	63.58	83.53	31.26	94.51	74.67	62.95	87.01	73.36	69.26	75.71	73.42
Clip-Adapter		92.49	65.96	84.43	32.10	93.90	78.25	63.59	87.84	74.01	69.55	76.76	74.44
Tip-Adapter-F		92.63	66.94	84.94	35.86	94.23	78.11	65.44	88.18	75.75	71.00	79.03	75.65
TaskRes		92.90	67.57	82.57	33.73	96.10	78.23	64.75	88.10	74.93	70.30	76.87	75.10
GraphAdapter		93.33	67.57	85.27	**36.87**	96.23	78.63	65.70	88.57	76.23	71.20	78.80	76.22
MKGPL (Ours)		**93.47**	**68.73**	**85.57**	36.57	**96.33**	**79.27**	**65.97**	**89.37**	**76.83**	**71.36**	**79.43**	**76.63**

datasets. These results collectively demonstrate that our proposed approach effectively leverages the positive and negative structural knowledge of nodes/classes for enhanced generalization capability. Moreover, as the number of shots increases, a noticeable improvement in the performance of all methods can be observed. Specifically, for our MKGPL, as the number of samples available for constructing the pre-built subgraphs increases (*i.e.*, the number of shots rises), the node features in the graph structure become increasingly accurate, leading to better performance.

Generalization. To evaluate the generalization capability of our proposed method, we conducted comprehensive experiments on four widely-used cross-domain benchmarks, including ImageNet-V2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. Experimental validation was performed across diverse visual backbone architectures including pre-trained ResNet-101, ViT-B/32, and ViT-B/16 models. As demonstrated in Table 2, our approach achieves superior generalization performance across all four domain-shifted datasets. Notably, when implemented with ResNet-50 backbone, our method outperforms GraphAdapter by 0.17 % in absolute accuracy. A critical distinction lies in parameterization requirements: while GraphAdapter necessitates increasing base text feature weights to 0.8 for optimal generalization performance, our method maintains consistent hyperparameter settings throughout both training and few-shot evaluation protocols. This observation underscores the methodological advantage of our framework: the integration of negative textual information subgraphs enables more precise boundary discrimination between positive and negative textual features. Consequently, the resulting learned representations exhibit enhanced expressiveness, leading to improved generalization capabilities across varied domain distributions and model architectures without requiring reset task-specific hyperparameters (*e.g.*, the base text feature weights).

4.3. Ablation studies

Effectiveness of different components. To validate the effectiveness of our individual components, we conducted ablation experiments under a 16-shot setup on 11 commonly used few-shot datasets. The experimental results are presented in Table 3. “Base” indicates that no graph structures or graph neural networks are utilized at all. Instead, it solely employs the original, frozen visual and text encoders to encode images and texts, and directly determines the category of the input image by using similarity metrics. “Ours (only P-N)” refers to the method incorporating image subgraph, positive/negative text subgraphs and the interactions between them, along with the integration of triplet loss item, but without using learnable Graph Prompts (GP). “Ours (only GP)” denotes the method that utilizes positive text and visual subgraphs and employs the corresponding GP, while it abandons the entire negative text subgraph and the triplet loss term. “Ours (P-N & GP)” extends graph prompt features to fully augment all three knowledge graphs and employs triplet loss for interaction. Different from the above configurations, MKGPL only insert learnable GP into positive text subgraph and image-specific subgraph. Prior studies [32,33] theoretically demonstrate that in graph convolutional networks with linear aggregation (*e.g.*, GCN), generic GP information can be broadly effective across multiple tasks. Our experiments also show that after introducing the negative text subgraph and performing convolution on structured knowledge from the tri-subgraph

Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

Table 2
Comparison of generalization among different methods using common backbones as the visual encoder. All methods listed were evaluated under the 16-shot experimental setting. ImageNet served as the training set, while various target datasets were used for testing to assess performance across different backbones under cross-domain conditions. (**Bold** and underlined values indicate the best and suboptimal performance, respectively).

Methods	Backbone	Source	Target				
		ImageNet	-V2	-Sketch	-A	-R	Avg.
Zero-shot CLIP	ResNet-50	58.18	51.34	33.32	21.65	56.00	40.58
Linear Probe CLIP		55.87	45.97	19.07	12.74	28.16	28.16
CoOp		62.95	55.11	32.74	22.12	54.96	41.23
TaskRes		64.75	56.47	35.83	22.80	60.70	43.95
GraphAdapter		65.70	56.40	34.50	21.88	58.94	42.93
GraphAdapter$_g$		64.94	56.58	35.89	23.07	**60.86**	44.10
MKGPL (Ours)		**65.97**	**56.89**	**36.01**	**23.37**	60.80	**44.27**
Zero-shot CLIP	ResNet-101	61.62	54.81	38.71	28.05	64.38	46.49
Linear Probe CLIP		59.75	50.05	26.80	19.44	47.19	35.87
CoOp		66.60	58.66	39.08	28.89	63.00	47.41
TaskRes		67.70	59.50	41.70	29.87	68.07	49.79
GraphAdapter		68.23	59.60	40.83	28.77	67.13	49.08
GraphAdapter$_g$		67.87	59.50	41.60	30.00	68.10	49.80
MKGPL (Ours)		**68.44**	**59.94**	**41.78**	**30.52**	**68.27**	**50.13**
Zero-shot CLIP	ViT-B/32	62.05	54.79	40.82	29.57	65.99	47.79
Linear Probe CLIP		59.58	49.73	28.06	19.67	47.20	36.17
CoOp		66.85	58.08	40.44	30.62	64.45	48.40
TaskRes		68.20	59.20	42.50	31.43	69.33	50.62
GraphAdapter		68.80	59.00	41.70	29.57	68.67	49.74
GraphAdapter$_g$		68.47	59.10	**42.70**	31.73	69.43	50.74
MKGPL (Ours)		**69.23**	**59.29**	42.60	**32.59**	**69.55**	**51.01**
Zero-shot CLIP	ViT-B/16	66.73	60.83	46.15	47.77	73.96	57.18
Linear Probe CLIP		65.85	56.26	34.77	35.68	58.43	46.29
CoOp		71.92	64.18	46.71	48.41	74.32	58.41
TaskRes		73.07	65.30	49.13	50.37	77.70	60.63
GraphAdapter		73.68	65.57	48.57	49.23	77.20	60.14
GraphAdapter$_g$		73.40	65.60	49.23	50.57	77.73	60.78
PiNI		71.74	64.40	48.25	48.57	74.39	58.90
MKGPL (Ours)		**73.80**	**65.74**	**49.49**	**50.65**	**77.83**	**60.93**

Table 3
Ablation study on different components of MKGPL under 16-shot setting. (**Bold** and underline values in the table represent the optimal and suboptimal performance, respectively).

Method	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
CLIP-Adapter	63.59	92.49	87.84	74.01	93.90	78.25	32.10	69.55	65.96	84.43	76.76	74.44
CoOp	62.95	91.83	87.01	73.36	94.51	74.67	31.26	69.26	63.58	83.53	75.71	73.42
TaskRes	64.75	92.90	88.10	74.93	96.10	78.23	33.73	70.30	67.57	82.57	76.87	75.10
Tip-Adapter	65.44	92.63	88.18	75.75	94.23	78.11	35.86	71.00	66.94	84.94	79.03	75.65
GraphAdapter	65.70	93.33	88.57	76.23	96.23	78.63	**36.87**	71.20	67.57	85.27	78.80	76.22
Base	58.18	86.29	85.77	55.61	66.14	77.31	17.28	58.52	42.32	37.56	61.46	58.77
Ours (only P-N)	65.83	93.20	88.63	76.67	95.93	79.17	35.77	71.23	**68.87**	85.47	79.27	76.37
Ours (only GP)	65.79	93.13	88.90	76.53	95.90	79.03	36.47	71.16	68.80	85.24	79.33	76.39
Ours (P-N & GP)	65.77	93.17	89.27	76.73	96.03	78.97	36.10	71.27	68.37	85.57	**79.57**	76.44
MKGPL	**65.97**	**93.47**	**89.37**	**76.83**	**96.33**	**79.27**	36.57	**71.36**	68.73	**85.57**	79.43	**76.63**

Table 4
Sensitivity analysis of different graph neural networks under 1/2/4/8/16-shot settings. (**Bold** values denote the optimal performance).

Method	Caltech101					DTD					Food101				
	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
Ours + GAT	88.47	88.90	90.67	91.73	92.87	**51.30**	54.78	58.17	64.20	68.37	74.37	74.50	77.84	78.17	78.90
Ours + GIN	65.47	76.57	82.06	85.35	90.10	37.73	45.03	53.77	61.17	66.67	41.23	52.50	57.83	66.50	68.77
Ours + GraphSAGE	86.00	88.27	88.96	89.32	91.83	50.04	53.45	58.87	63.60	66.90	66.87	67.73	68.70	71.83	74.53
MKGPL	**89.20**	**90.37**	**91.73**	**92.35**	**93.47**	51.17	**55.18**	**59.77**	**64.70**	**68.73**	**77.40**	**77.47**	**78.34**	**78.53**	**79.27**

framework, our method achieves an average performance improvement of 0.15 % compared to the SOTA method (*i.e.*, GraphAdapter). By incorporating graph node prompts into text and visual subgraphs, we observe a 0.17 % performance gain over GraphAdapter. This demonstrates that both negative text subgraph integration and graph prompt features contribute positively to few-shot classification tasks. Notably, fully applying graph prompt features to all three knowledge graphs yields only marginal improvements compared to using them exclusively for positive text and visual subgraphs. Therefore, our final methodology employs graph prompt features for positive text and visual

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله
مورد تایید ایران پیپر

Table 5
Sensitivity analysis of GCN depth under 8/16-shot settings. (**Bold** values denote the optimal performance).

GCN Layers	OxfordPets		DTD		Caltech101	
	8	16	8	16	8	16
1 (Ours)	**87.93**	**89.37**	**64.70**	**68.37**	**92.35**	**93.47**
2	85.43	87.50	64.00	67.93	91.29	92.63
3	82.23	84.50	63.26	67.53	90.17	92.33
4	79.76	84.03	62.53	67.53	89.82	92.20

subgraphs, while their application to negative text subgraphs remains optional.

Sensitivity analysis of different GNNs. To probe the impact of graph neural network (GNN [20]) architectures on performance of MKGPL, we replaced GCN with other GNNs like GraphSAGE [21], GAT [22], and GIN [25] for experiments. Comparative results are in Table 4. In most settings, MKGPL with GCN outperforms others, showing stronger robustness and generalization. This may stem from the inherent traits of these GNNs. GAT computes dynamic connection weights via node feature attention, with edge weights as node-driven dynamic coefficients, suitable for heterogeneous and dynamic graphs. GIN uses multi-layer perceptrons for feature extraction, excelling at graph isomorphism distinction and complex structure modeling, with feature discrimination ensured by the Weisfeiler-Lehman test. GraphSAGE employs flexible neighbor sampling and aggregation, handling unseen node features without full graph infomations, ideal for large-scale graphs. GCN approximates first-order neighbors via ChebNet, aggregating neighbor features with predefined weights and then first-order ones based on topological edge weights. Our experiments feature small-scale graph data, closely and uniformly connected nodes (*i.e.*, no big disparity in central-neighbor connection tightness), so GCN excels. These findings show our framework works with various GNNs. But for few-shot learning, GCN offers the best balance between structural representation and training stability.

Sensitivity analysis of hyperparameters α and β. To evaluate the influence of hyperparameter configurations on our method's overall performance, we conducted ablation studies on hyperparameters α and β. Here, α controls the weighting between positive text subgraph and image-specific subgraph, while β balances the raw text features and the graph convolutional fused features. As shown in Fig. 4, our method achieves optimal performance when α and β are set to 0.7 and 0.6, respectively. Additionally, we design an experiment where α and β were implemented as learnable parameters (denoted as “learnable”). The results indicate that making these parameters learnable yields inferior performance compared to fixing their values. This performance degradation is likely attributed to the challenges of effectively optimizing α and β in few-shot scenarios, where limited training samples hinder the self-learning process of these parameters. These findings suggest that while adaptive parameter tuning holds theoretical appeal, manual calibration of α and β provides more reliable performance guarantees for our framework under data-constrained conditions. The optimal fixed values identified in this study are therefore adopted as the default configuration in our final implementation.

Sensitivity analysis of margin m and balance weight γ. To investigate the impact of different values of the margin m in the triplet loss (refer to Eq. (5)) and varying balance weight γ in the final loss $\mathcal{L}$ on the performance of the MKGPL, we designed corresponding experiments. The results are illustrated in Fig. 5(a)(b) and (c)(d), respectively. These experiments were conducted on the Caltech101 and EuroSAT. Through the experimental curves, it can be observed that, irrespective of the number of shots, setting both the margin m and γ to 1.0 maximizes the overall model performance. Moreover, MKGPL exhibits relatively stable performance under different values of m or γ. This further demonstrates that our proposed method displays relatively consistent performance and robustness across different shot number settings.

Table 6
Model complexity comparison on ImageNet.

Method	Training time/epoch (s)	Epochs	Parameters (M)	Performance
CoOp	40.91	200	0.008	62.95
CLIP-Adapter	45.71	200	0.524	63.59
Tip-Adapter-F	12.36	20	16.380	65.51
GraphAdapter	23.29	150	4.145	65.70
MKGPL (Ours)	28.57	70	10.320	65.97

Sensitivity analysis of GCN layers. To investigate the influence of GCN layer depth on our proposed MKGPL, we conducted experiments using 1 to 4 GCN layers during the graph prompt learning phase. Performance evaluations were performed under 8-shot and 16-shot settings across 4 benchmark datasets, including OxfordPets, DTD, and Caltech101. As shown in Table 5, the model achieved optimal or near-optimal performance across all datasets when employing a single GCN layer. Notably, increasing the number of GCN layers generally led to performance degradation. This phenomenon may be attributed to the oversmoothing [20] issue in multi-layer GCNs, where node features tend to converge excessively, thereby weakening the discriminative capability between different classes. This effect becomes particularly pronounced in few-shot scenarios. Consequently, maintaining an appropriate GCN layer depth is critical for preserving semantic discriminability, with the single-layer GCN configuration demonstrating superior performance in our framework.

Strong robustness across different backbones. To evaluate the efficacy of MKGPL across diverse CLIP visual backbones, we conducted experimental evaluations to assess the performance when utilizing different backbones as visual encoders. Specifically, we employed 4 representative architectures, including ResNet-50, ResNet-101, ViT-B/32 [51], and ViT-B/16. As demonstrated in Fig. 6, our method MKGPL consistently outperforms previous approaches across all backbone variants. This indicates that MKGPL exhibits remarkable insensitivity to architectural variations with strong robustness, achieving SOTA performance regardless of whether the visual backbone employs ViT or ResNet architectures.

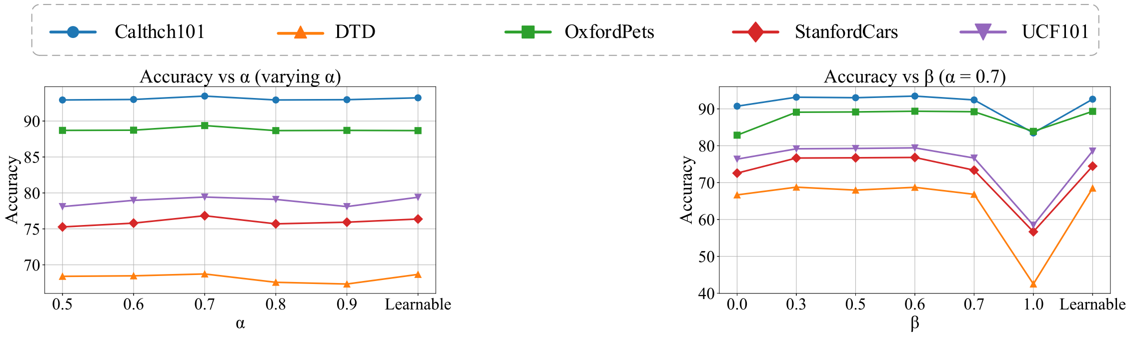
4.4. *Model complexity analysis*

To evaluate the model complexity of MKGPL, we conduct experiments on the large-scale ImageNet dataset under the 16-shot setting, reporting the computational complexity, training time, and learnable parameter scale in Table 6. It is observed that our method achieves SOTA accuracy performance. While exhibiting a marginal increase in parameter count and training time compared to the GraphAdapter, our approach demonstrates reduced training epochs. This phenomenon may be attributed to the introduction of a subgraph configuration and additional learnable graph prompt vectors. Notably, when compared to Tip-Adapter-F, our method not only improves performance but also reduces the total learnable parameters. Furthermore, significant reductions in both training time and epochs are achieved relative to CoOp and CLIP-Adapter. These experimental results indicate that while maintaining superior performance, our method only introduces minimal additional parameters and time overhead compared to existing approaches.

4.5. *Visualization*

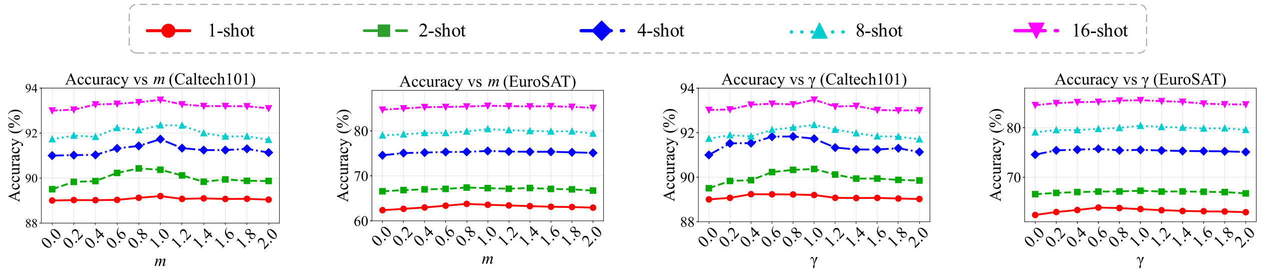
t-SNE visualization. On the one hand, to demonstrate the superiority of the proposed MKGPL in prompt tuning for VLMs, we visualize the changes in data samples representing textual features before and after graph structure intervention. As illustrated in Fig. 7(a), 20 classes were randomly sampled from the Food101 [46], and the t-SNE is employed to visualize the distribution of each graph node in classification tasks, where each graph node represents the position of a class of samples in the latent space. Observations reveal that, compared to the “Base”

Downloaded from https://iranpaper.ir https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر



(a) Different values of α (b) Different values of β

Fig. 4. Sensitivity analysis of different hyperparameters α, β.



(a) Caltech101 (b) EuroSAT (c) Caltech101 (d) EuroSAT

Fig. 5. Sensitivity analysis of different values of margin m and balance weight γ under 1-/2-/4-/8-/16-shot settings on Caltech101 and EuroSAT. (a)(b) Performance with different margin m. (c)(d) Performance with various balance weight γ.

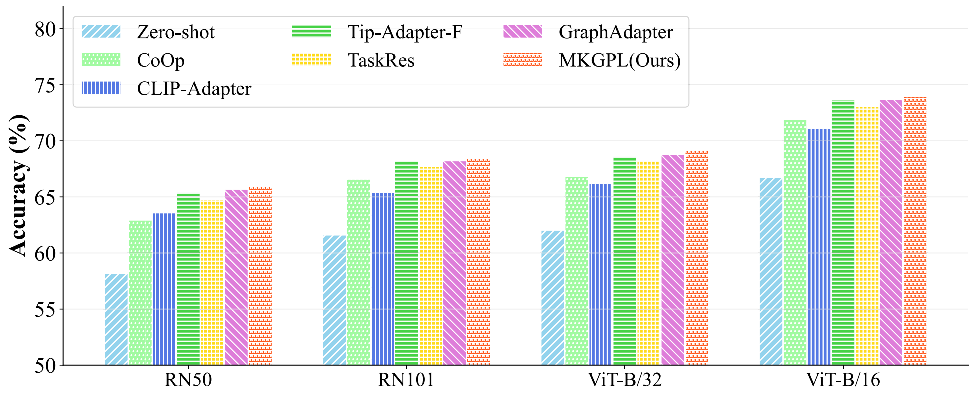


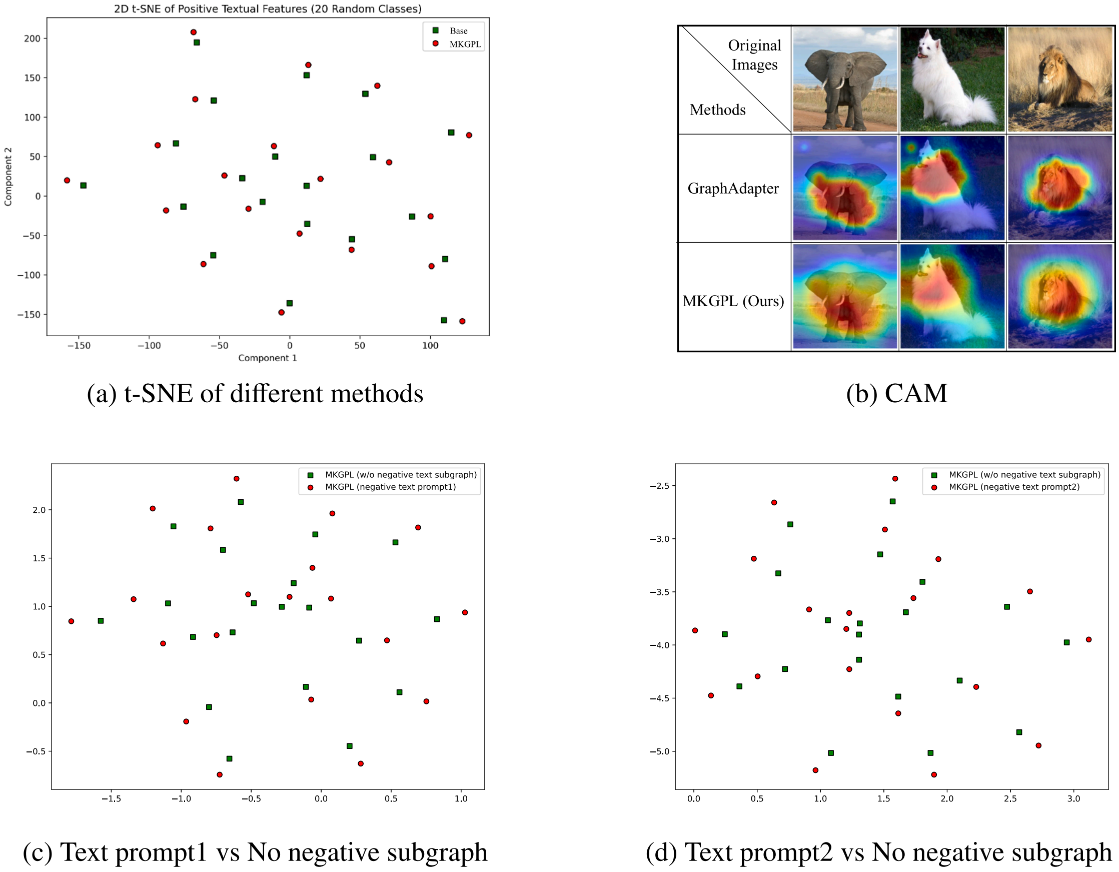
Fig. 6. Comparisons of different backbones on ImageNet under 16-shot setting.

scheme, MKGPL could position nodes of different classes farther apart in the latent space, thereby increasing inter-class distances compared to the “Base”. This indicates that MKGPL optimization effectively enhances the performance of prompt-based tuning for VLMs and subsequently improves classification performance.

On the other hand, to illustrate that the utilization of different negative text prompts can enhance the performance of MKGPL, we follow the aforementioned sampling approach and present the t-SNE visualization results of MKGPL without using negative text subgraph at all and with two different negative text prompts (*i.e.*, “text prompt1” is “This is NOT a photo of”, “text prompt2” is “The image does NOT include the class of”), corresponding to Fig. 7(c) and (d), respectively. From the two t-SNE maps, it can be seen that using different negative textual prompts enables different nodes to be located at relatively farther positions in the common space, indicating the effectiveness of the negative text branch and negative text prompts employed in MKGPL.

CAM visualization. To intuitively demonstrate the effectiveness of MKGPL in image semantic region localization, Fig. 7(b) presents visual comparisons of Class Activation Map (CAM) between different approaches. We contrast the attention distribution of MKGPL with the prior GraphAdapter across multiple images. As we see, the first, second, and third row respectively showcase randomly selected original images, CAMs from GraphAdapter, and CAMs from MKGPL. Compared to GraphAdapter, our approach demonstrates superior capability in precisely focusing on discriminative regions of target objects. For instance, in the elephant image, our method significantly concentrates

Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر



(a) t-SNE of different methods

(b) CAM

(c) Text prompt1 vs No negative subgraph

(d) Text prompt2 vs No negative subgraph

Fig. 7. (a) Visualization of graph node variations before and after MKGPL application. Each node corresponds to the representation of a specific class. (b) CAM visualization comparison between our MKGPL and GraphAdapter. (c)(d) t-SNE visualization comparison between the scenario where MKGPL incorporates text prompt1/prompt2 and the scenario where MKGPL completely excludes the negative text subgraph.

attention on the body area. Similarly, for images of white dogs and lions, our CAMs better cover their primary body regions. In contrast, GraphAdapter exhibits diffused attention or background deviation in certain samples, while MKGPL generates more concentrated and discriminative attention regions.

5. Discussions

Limitations. Although MKGPL has achieved favorable performance in few-shot learning tasks, it still has certain limitations. On the one hand, since an independent tri-subgraph structure needs to be constructed for each dataset, this hinders the model's generality and flexibility while increasing the size of the model files. On the other hand, for large-scale image datasets (such as ImageNet), the graph construction process often faces the risk of memory explosion, leading to the current approach of using batch-wise graph construction. This may impede the efficient training of the whole model. In practical applications, batch-wise construction can be employed to build larger sub-graph structures within limited memory space. Moreover, techniques such as matrix factorization can be considered to handle excessively large adjacency matrices. In summary, we will strive to further address these limitations in the future. **Future work.** In the future, we will conduct research in the fields of few-shot learning or cross-modal learning, focusing on aspects such as dynamic graph structures [52,53], personalized information prompts, and cross-modal graph prompts [54]. Specifically, dynamic graph structures will facilitate the construction of dataset-independent graph structures. Personalized information prompts can help leverage lightweight data to express fine-grained differences in information across different data distributions. Cross-modal graph prompts will enable smoother information interaction between images and text without significantly increasing computational and storage overheads, thereby contributing to the performance enhancement in few-shot learning scenarios.

6. Conclusion

This paper proposes Graph Prompt Learning with Multi-view Knowledge (*abbr.*, MKGPL) framework that synergistically integrates positive/negative textual descriptions with visual information through prompt learning. The proposed approach constructs dedicated subgraph structures for positive textual features, negative textual features, and visual features to capture structured knowledge representations. During inference, testing textual features are dynamically incorporated as new nodes into each subgraph, followed by the insertion of learnable prompt vectors to enhance model generalization and robustness. The framework enables bidirectional information fusion through multi-view parallel interactions, *i.e.*, (1) intra-modal text-text alignment between the testing text with positive/negative subgraphs, (2) cross-modal text-image fusion through the testing text with visual subgraph, and (3) adaptive knowledge aggregation facilitated by prompt-enhanced graph convolutions. Comprehensive experimental evaluations on 11 datasets demonstrate the superior performance of MKGPL over SOTA baselines across multiple benchmarks.

CRediT authorship contribution statement

Yanzhao Xie: Conceptualization, Investigation, Writing – original draft; **Man Qiu:** Conceptualization, Validation; **Yangtao Wang:** Conceptualization, Writing – review & editing; **Siyuan Chen:** Conceptualization, Writing – review & editing; **Meie Fang:** Conceptualization,

iranpaper Downloaded from https://iranpaper.ir

https://www.tarjomano.com ترجمه تخصصی این مقاله مورد تایید ایران پیپر

Supervision; **Maobin Tang:** Conceptualization, Supervision; **Wensheng Zhang:** Conceptualization, Supervision.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62506085, No. 62406082), Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515110659, No. 2023A1515110650), Guangzhou Specialized Program for Basic and Applied Basic Research (No. 2025A04J4590), Guangzhou Science and Technology Planning Project (No. 2024A03J0013), Guangdong Provincial Department of Education Innovation Strong School Youth Innovation Talent Project (No. 2023KQNCX055), Science and Technology Program of Guangzhou (No. SL2023A03J00284).

References

[1] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: open and efficient foundation language models, CoRR abs/2302.13971 (2023).

[2] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, ICML, 139 of *Proceedings of Machine Learning Research*, 2021, pp. 8748–8763.

[3] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Learning to prompt for vision-language models, Int. J. Comput. Vis. 130 (9) (2022) 2337–2348.

[4] K. Zhou, J. Yang, C.C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, IEEE, 2022, pp. 16795–16804.

[5] T. Yu, Z. Lu, X. Jin, Z. Chen, X. Wang, Task residual for tuning vision-language models, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, 2023, pp. 10899–10909.

[6] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, H. Li, Tip-adapter: training-free adaption of CLIP for few-shot classification, in: Computer Vision - ECCV 2022 - 17th European Conference, 13695 of *Lecture Notes in Computer Science*, 2022, pp. 493–510.

[7] F. Zhang, R. Wei, Y. Xie, Y. Wang, X. Tan, L. Ma, M. Tang, L. Fan, Cross-coupled prompt learning for few-shot image recognition, Displays 85 (2024) 102862.

[8] S. Zhang, W. Luo, D. Cheng, Y. Xing, G. Liang, P. Wang, Y. Zhang, Prompt-based modality alignment for effective multi-modal object re-identification, IEEE Trans. Image Process. 34 (2025) 2450–2462.

[9] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, Y. Qiao, CLIP-adapter: better vision-language models with feature adapters, Int. J. Comput. Vis. 132 (2) (2024) 581–595.

[10] X. Li, D. Lian, Z. Lu, J. Bai, Z. Chen, X. Wang, GraphAdapter: tuning vision-language models with dual knowledge graph, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023.

[11] F. Tao, G. Xie, F. Zhao, X. Shu, Kernel-aware graph prompt learning for few-shot anomaly detection, in: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, 2025, pp. 7347–7355.

[12] T. Fang, Y. Zhang, Y. Yang, C. Wang, L. Chen, Universal prompt tuning for graph neural networks, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023.

[13] F. Zhou, C. Cao, Overcoming catastrophic forgetting in graph neural networks with experience replay, in: AAAI Conference on Artificial Intelligence, AAAI 2021, 2021, pp. 4714–4722.

[14] J. Jinghong, L. Song, J. Li, Y. Kong, HePa: heterogeneous graph prompting for all-level classification tasks, in: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, 2025, pp. 11915–11923.

[15] C. Liu, J. Wen, Y. Xu, B. Zhang, L. Nie, M. Zhang, Reliable representation learning for incomplete multi-view missing multi-label classification, IEEE Trans. Pattern Anal. Mach. Intell. 47 (6) (2025) 4940–4956.

[16] M. Xu, J. Wang, M. Tao, B. Bao, C. Xu, CookGALIP: recipe controllable generative adversarial CLIPs with sequential ingredient prompts for food image generation, IEEE Trans Multimed. 27 (2025) 2772–2782.

[17] L. Chen, X. Wang, J. Lu, S. Lin, C. Wang, G. He, CLIP-driven open-vocabulary 3D scene graph generation via cross-modality contrastive learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, 2024, pp. 27863–27873.

[18] P. Mangai, M.K. Geetha, G. Kumaravelan, Two-stream spatial-temporal feature extraction and classification model for anomaly event detection using hybrid deep learning architectures, Int. J. Image Graph. 24 (06) (2024) 2450052.

[19] J. Wen, J. Long, X. Lu, C. Liu, X. Fang, Y. Xu, Partial multiview incomplete multilabel learning via uncertainty-driven reliable dynamic fusion, IEEE Trans. Pattern Anal. Mach. Intell. (2025).

[20] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, ICLR 2017, 2017.

[21] W.L. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Advances in Neural Information Processing Systems, 2017, pp. 1024–1034.

[22] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: 6th International Conference on Learning Representations, ICLR 2018, 2018.

[23] S. Wang, X. Liu, Q. Liao, Y. Wen, E. Zhu, K. He, Scalable multi-view graph clustering with cross-view corresponding anchor alignment, IEEE Trans. Knowl. Data Eng. 37 (5) (2025) 2932–2945.

[24] J. Wen, G. Xu, Z. Tang, W. Wang, L. Fei, Y. Xu, Graph regularized and feature aware matrix factorization for robust incomplete multi-view clustering, IEEE Transact. Circuit. Syst. Video Technol. 34 (5) (2024) 3728–3741.

[25] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, in: 7th International Conference on Learning Representations, ICLR 2019, 2019.

[26] R. Chen, T. Chen, X. Hui, H. Wu, G. Li, L. Lin, Knowledge graph transfer network for few-shot recognition, in: The AAAI Conference on Artificial Intelligence, AAAI 2020, 2020, pp. 10575–10582.

[27] M. Liu, K. Liang, S. Wang, X. Hu, S. Zhou, X. Liu, Deep temporal graph clustering: a comprehensive benchmark and datasets, IEEE Trans. Pattern Anal. Mach. Intell. (2025) 1–18.

[28] Y. Lu, J. Tan, S. Zhang, Y. Xing, G. Liang, Y. Zhang, Nearest-neighbor class prototype prompt and simulated logits for continual learning, Pattern Recognit. 169 (2026) 111933.

[29] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (9) (2023) 195:1–195:35.

[30] Y. Fu, X. Zhu, X. Ji, Z. Tang, J. Wen, Weighted local-global facial feature-based VMamba network for Williams syndrome diagnosis, Int. J. Image Graph. (2025).

[31] M.U. Khattak, H.A. Rasheed, M. Maaz, S.H. Khan, F.S. Khan, MaPLe: multi-modal prompt learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 19113–19122.

[32] X. Sun, H. Cheng, J. Li, B. Liu, J. Guan, All in one: multi-task prompting for graph neural networks, in: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, 2023, pp. 2120–2131.

[33] Q. Wang, X. Sun, H. Cheng, Does graph prompt work? A data operation perspective with theoretical analysis, CoRR abs/2410.01635 (2024).

[34] X. Sun, J. Zhang, X. Wu, H. Cheng, Y. Xiong, J. Li, Graph prompt learning: a comprehensive survey and beyond, CoRR abs/2311.16534 (2023).

[35] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, ICLR 2017, OpenReview.net, 2017.

[36] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[37] J. Krause, M. Stark, J. Deng, L. Fei-Fei, 3D Object representations for fine-grained categorization, in: IEEE International Conference on Computer Vision Workshops, ICCV Workshops, 2013, pp. 554–561.

[38] K. Soomro, A.R. Zamir, M. Shah, UCF101: a dataset of 101 human actions classes from videos in the wild, CoRR abs/1212.0402 (2012).

[39] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, Comput. Vision Image Understand. 106 (1) (2007) 59–70.

[40] M. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP, 2008, pp. 722–729.

[41] J. Xiao, J. Hays, K.A. Ehinger, A. Oliva, A. Torralba, SUN Database: large-scale scene recognition from abbey to zoo, in: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, 2010, pp. 3485–3492.

[42] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi, Describing textures in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2014, pp. 3606–3613.

[43] P. Helber, B. Bischke, A. Dengel, D. Borth, EuroSAT: a novel dataset and deep learning benchmark for land use and land cover classification, IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 12 (7) (2019) 2217–2226.

[44] S. Maji, E. Rahtu, J. Kannala, M.B. Blaschko, A. Vedaldi, Fine-grained visual classification of aircraft, CoRR abs/1306.5151 (2013).

[45] O.M. Parkhi, A. Vedaldi, A. Zisserman, C.V. Jawahar, Cats and dogs, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, 2012, pp. 3498–3505.

[46] L. Bossard, M. Guillaumin, L.V. Gool, Food-101 - mining discriminative components with random forests, in: Computer Vision - ECCV 2014 - 13th European Conference, 8694 of *Lecture Notes in Computer Science*, 2014, pp. 446–461.

[47] B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do ImageNet classifiers generalize to ImageNet?, in: Proceedings of the 36th International Conference on Machine

iranpaper Downloaded from https://iranpaper.ir
https://www.tarjomano.com ترجمه تخصصی این مقاله
مورد تایید ایران پیپر

Learning, ICML 2019, 97 of *Proceedings of Machine Learning Research*, 2019, pp. 5389–5400.

[48] H. Wang, S. Ge, Z.C. Lipton, E.P. Xing, Learning robust global representations by penalizing local predictive power, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 2019, pp. 10506–10518.

[49] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 97 of *Proceedings of Machine Learning Research*, 2019, pp. 2790–2799.

[50] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, J. Gilmer, The many faces of robustness: a critical analysis of out-of-distribution generalization, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, 2021, pp. 8320–8329.

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, 2021.

[52] Z. Shu, X. Sun, H. Cheng, When LLM meets hypergraph: a sociological analysis on personality via online social networks, in: Proceedings of the ACM International Conference on Information and Knowledge Management, CIKM 2024, 2024, pp. 2087–2096.

[53] X. Sun, H. Yin, B. Liu, Q. Meng, J. Cao, A. Zhou, H. Chen, Structure learning via meta-hyperedge for dynamic rumor detection, IEEE Trans. Knowl. Data Eng. 35 (9) (2023) 9128–9139.

[54] H. Zhang, C. Shen, X. Sun, J. Tan, Y. Rong, C. Piao, H. Cheng, L. Yi, Adaptive coordinators and prompts on heterogeneous graphs for cross-domain recommendations, CoRR abs/2410.11719 (2024).

Yanzhao Xie currently serves as an associate professor at the School of Computer Science and Cyber Engineering, Guangzhou University, China. He received his Ph.D. degree from Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China, in 2023. His main research interests include multimodal retrieval, image recognition, graph neural network, multimedia content analysis and reinforcement learning. He has published papers in international conferences and journals including IJCAI, CIKM, ICMR, APWeb-WAIM, TMM, etc.

Man Qiu is currently pursuing the master's degree at the School of Computer Science and Cyber Engineering Guangzhou University, Guangzhou, China, under the supervision of associate professor Maobin Tang. His current research interests include computer vision, few-shot learning, etc.

Yangtao Wang currently serves as an associate professor at the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. He received the Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China, in 2021. His main research interests include computer vision, multi-modal representation, and graph neural networks. He has published more than 30 papers in international conferences and journals including AAAI, IJCAI, CIKM, ICMR, ICME, WSDM, World Wide Web Journal, Pattern Recognition, IEEE Transactions on Multimedia, ACM Transactions on Data Science, etc.

Siyuan Chen currently serves as an associate professor at the School of Computer Science and Cyber Engineering, Guangzhou University, China. He received the Ph.D. degree in computer science and technology from Sun Yat-sen University, Guangzhou, China in 2023. His research interests include graph learning, spatio-temporal data mining, and urban computing.

Meie Fang received the Ph.D. degree in applied mathematics from Zhejiang University, Hangzhou, China. She is currently a Full Professor with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. She worked in the Institute of Computer Graphics and Image, Hangzhou Dianzi University from June 2007 to June 2017, and was transferred to Guangzhou University in June 2017. She has served as postdoctoral in the State Key Lab of CAD & CG, Zhejiang University and the Postdoctoral Station of Computer Application Technology, Shanghai Jiao Tong University. She visited City University of Hong Kong and Purdue University of the United States for the purpose of academic exchange several times in recent years. Her current research interests include intelligent computer graphics, medical image analysis, and AI security.

Maobin Tang is currently an associate professor at the School of Computer Science and Cyber Engineering, Guangzhou University, China. He received the Ph.D. degree from South China Agricultural University, Guangzhou, China. His main research interests include artificial intelligence, data mining, system analysis and design. He has published more than 20 papers and won the first prize of “Excellent Teaching Achievements of the university” and “Excellent Teaching Achievements of Guangdong Province”.

Wensheng Zhang received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, in 2000. He joined the Institute of Software, CAS, in 2001. He is currently a Professor in machine learning and data mining and the Director of the Research and Development Department, Institute of Automation, CAS. He is also with Guangzhou University, Guangzhou, China. His research interests include computer vision, pattern recognition, and artificial intelligence.