GCT: Graph Co-Training for Semi-Supervised Few-Shot Learning

Rui Xu, Lei Xing, Shuai Shao, *Member, IEEE*, Lifei Zhao, Baodi Liu, *Member, IEEE*, Weifeng Liu, *Senior Member, IEEE*, and Yicong Zhou, *Senior Member, IEEE*

***Abstract*— Few-shot learning (FSL), purposing to resolve the problem of data-scarce, has attracted considerable attention in recent years. A popular FSL framework contains two phases: (i) the pre-train phase employs the base data to train a CNN-based feature extractor. (ii) the meta-test phase applies the frozen feature extractor to novel data (novel data has different categories from base data) and designs a classifier for recognition. To correct few-shot data distribution, researchers propose Semi-Supervised Few-Shot Learning (SSFSL) by introducing unlabeled data. Although SSFSL has been proved to achieve outstanding performances in the FSL community, there still exists a fundamental problem: the pre-trained feature extractor cannot adapt to the novel data flawlessly due to the cross-category setting. Usually, large amounts of noises are introduced to the novel feature. We dub it as Feature-Extractor-Maladaptive (FEM) problem. To tackle FEM, we make two efforts in this paper. First, we propose a novel label prediction method, Isolated Graph Learning (IGL). IGL introduces the Laplacian operator to encode the raw data to graph space, which helps reduce the dependence on features when classifying, and then project graph representation to label space for prediction. The key point is that: IGL can weaken the negative influence of noise from the feature representation perspective, and is also flexible to independently complete training and testing procedures, which is suitable for SSFSL. Second, we propose Graph Co-Training (GCT) to tackle this challenge from a multi-modal fusion perspective by extending the proposed IGL to the co-training framework. GCT is a semi-supervised method that exploits the unlabeled samples with two modal features to crossly strengthen the IGL classifier. We estimate our method on five benchmark few-shot learning datasets and achieve outstanding performances compared with other state-of-the-art methods. It demonstrates the effectiveness of our GCT.**

***Index Terms*— Few-shot learning, semi-supervised few-shot learning (SSFSL), feature-extractor-maladaptive (FEM), isolated graph learning (IGL), graph co-training (GCT).**

I. Introduction

In RECENT years, the performance of computer vision tasks based on deep learning has reached or even surpassed the human beings' level, such as image classification [1]–[3], person re-identification [4]–[6], and point cloud recognition [7]–[9]. The adequate labeled data plays a crucial role for the success. However, it is a challenge for data collection and maintenance in real-world situations. To this end, few-shot learning (FSL), as a pioneer work to address the lack of labeled samples for each category, has aroused widespread concerns.

In a standard FSL, the employed data includes two parts, *i.e.*, base set and novel set. There are many labeled samples in the base set, but very few in the novel set (typically, for the general FSL setting, each category only has 1 or 5 labeled samples). Notably, the categories contained in the base set are entirely different from those in the novel set. Generally, researchers split the FSL model into two phases: (i) pre-train. Training a feature extractor through the base set. (ii) meta-test. First, employing the feature extractor to extract the features of novel data, and then designing a classifier to recognize the novel data's category. Besides, to overcome overfitting caused by the *few-shot* setting, researchers prefer to decouple the complete model, that is, freezing the parameters of the feature extractor after pre-training and directly extracting the cross-category novel features in the meta-test phase.

During the stage of designing the classifier, the FSL-based methods can be categorized into two sorts according to the type of data employed: (i) supervised FSL, and (ii) semi-supervised FSL. Specifically, the novel data includes three components: support data (*i.e.*, labeled training data), unlabeled data (*i.e.*, unlabeled training data), and query data (*i.e.*, to-be-classified testing data). The difference between the two

Manuscript received 30 April 2022; revised 13 July 2022; accepted 27 July 2022. Date of publication 4 August 2022; date of current version 6 December 2022. This work was supported in part by the Yunnan Key Laboratory of Media Convergence; in part by the Natural Science Foundation of Shandong Province, China, under Grant ZR2019MF073; in part by the Shandong Provincial Key Laboratory of Computer Network through the Open Research Fund (SDKLCN-2018-01); in part by the Qingdao Science and Technology Project (17-1-1-8-jch); in part by the Fundamental Research Funds for the Central Universities, China University of Petroleum, China, under Grant 20CX05001A; in part by the Major Scientific and Technological Projects of the China National Petroleum Corporation (CNPC) under Grant ZD2019-183-008; in part by the Creative Research Team of Young Scholars at Universities in Shandong Province under Grant 2019KJN019; and in part by the Graduate Innovation Project of China University of Petroleum under Grant YCX2021123 and Grant YCX2021117. This article was recommended by Associate Editor S. Wang. *(Rui Xu and Lei Xing are co-first authors.) (Corresponding authors: Baodi Liu; Weifeng Liu.)*

Rui Xu is with the College of Control Science and Engineering, China University of Petroleum, Qingdao 266580, China, and also with the Yunnan Key Laboratory of Media Convergence, Kunming 650000, China.

Lei Xing and Lifei Zhao are with the College of Oceanography and Space Informatics, China University of Petroleum, Qingdao 266580, China.

Shuai Shao is with the College of Control Science and Engineering, China University of Petroleum, Qingdao 266580, China, and also with the Zhejiang Laboratory, Hangzhou 311100, China.

Baodi Liu and Weifeng Liu are with the College of Control Science and Engineering, China University of Petroleum, Qingdao 266580, China (e-mail: thu.liubaodi@gmail.com; liuwf@upc.edu.cn).

Yicong Zhou is with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau 999078, China.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3196550.

Digital Object Identifier 10.1109/TCSVT.2022.3196550

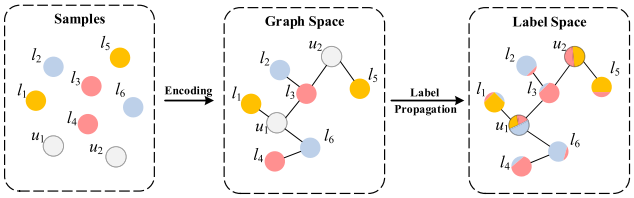


Fig. 1. Isolated graph learning (IGL) classifier. $l.$ and $u.$ denote the labeled and unlabeled samples, respectively. IGL first encodes the samples to graph representation and then propagate the label information through graph structure for prediction.

settings is whether to use the unlabeled data when building the classifier. For more details, please see Section III.

Compared with the supervised FSL, semi-supervised approaches [10]–[13] can effectively correct few-shot data distributions to make the learned classifiers have higher quality. However, to achieve this goal, it must be on the premise that sample features are less noisy. But unfortunately, a fundamental problem in FSL, Feature-Extractor-Maladaptive (FEM), is easy to break up the assumption. Specifically, researchers obtain a feature extractor in the pre-train process and apply it directly to the meta-test process, which is challenging to ensure that the frozen feature extractor is capable of adapting to the novel categories. To solve this challenge, we make two efforts in this paper.

First, inspired by [14], we know that transforming the raw data to graph representation is helpful in reducing the dependence on features in classification tasks. To this end, we propose a novel label prediction method dubbed as Isolated Graph Learning (IGL), try to weaken the negative impact of noise from the feature representation perspective. The framework of IGL is shown in Figure 1. We first introduce the graph Laplacian operator to encode the sample's feature embedding to graph space and then project the graph representation to label space for prediction by regularization. Compared with the traditional graph learning method [15], needing both labeled and unlabelled data to propagate label information, our IGL is more flexible to independently complete training and testing procedures. Furthermore, compared with graph neural network (GNN) based methods, there are no abundant parameters in our IGL that need to be updated with the propagation of deep neural networks. In other words, it is easy to be implemented. These attributes are very friendly for few-shot classification in the semi-supervised setting.

Second, we propose Graph Co-Training (GCT) to weaken the negative effect of noise from a multi-modal fusion perspective. Suppose that we have features from different feature extractors, we can integrate different predictions (obtained from different features) through collaborative training (co-training) to complete the final classification. To be more specific, we first try to get two-modal features from two designed feature extractors. In this phase, we have the flexibility to adapt the classical models in few-shot communities. This paper uses two kinds of self-supervision ways from [16], [17]. Then, we exploit the support data to train two basic classifiers (*i.e.*, IGL) with different modal features. Next, we separately predict the unlabeled data with the two modalities of classifiers. At last, we apply the unlabeled data with the most confident predictions to crossly update the classifiers. The designed co-training way is mainly to strengthen our classifier's robustness to reduce the interference caused by FEM. We illustrate the framework of our GCT in Figure 2. For convenience, we list some crucial abbreviations and notations in Table I.

In summary, the main contributions focus on:

- Aiming at the Feature-Extractor-Maladaptive (FEM) problem in semi-supervised few-shot learning, we propose a novel graph learning based classifier dubbed as Isolated Graph Learning (IGL), which completes training and testing procedures independently.
- We combine our IGL with the co-training framework to design a Graph Co-training (GCT) algorithm. It extends IGL to semi-supervised few-shot learning through fusing multi-modal information.
- The comparison results with SOTAs on five benchmark FSL datasets have evaluated the efficiency of our GCT.

II. Related Work

A. Semi-Supervised Few-Shot Learning

Recently, semi-supervised few-shot learning (SSFSL) has attracted lots of attention. Researchers assume that abundant unlabeled data is available to be used for constructing the classifier. They introduce various traditional semi-supervised learning methods to the few-shot learning (FSL) task. Here, we list several classical semi-supervised approaches and corresponding FSL works. (i) Consistency regularization methods aim to improve the robustness of classifiers when the images are noisy. MetaMix [18], BR-ProtoNet [19] *et.al.* promote the classifier from this way. (ii) Self-training methods first train a classifier with labeled data, then exploit it to generate pseudo labels for unlabeled data, and at last update the classifier with pseudo-labeled data. Recent self-training based FSL methods, including LST [10] ICI [11], PLCM [12], iLPC [13] have achieved outstanding performances. (iii) Hybrid based FSL methods, such as MixMatch [20], and FixMatch [21], try to construct a unified framework of semi-supervised learning by combining several current dominant approaches such as self-training and consistency regularization.

B. Multi-Modal Few-Shot Learning

As there are two sides to every coin, it is boundedness to define objects from a single point of view. Multi-view learning as an effective strategy has attracted extensive attention in the past decade. In few-shot learning, some similar methods have been proposed, such as: DenseCls [22], its feature map is divided into various blocks, and the corresponding labels are predicted; DivCoop [23] trains the feature extractors on various datasets and integrates them into a multi-domain representation; URT [24] is an improved method compared with DivCoop [23], which proposes a transformer layer to help the network employ various datasets; DWC [25] introduces a cooperate strategy on a designed ensemble model to integrate multiple information. Although the above-mentioned approaches are based on multi-modal learning, they are limited by the fused feature extractors and

TABLE I
THE DEFINITION OF ABBREVIATIONS AND NOTATIONS

Abbreviation and Notation	Definition
IGL	Isolated Graph Learning
GCT	Graph Co-Training
FEM	feature-extractor-maladaptive
Std-Mod	standard modality
Meta-Mod	meta modality
SS-R-Mod	self-supervised rotation modality
SS-M-Mod	self-supervised mirror modality
FSL	few-shot learning
ISFSL	inductive supervised few-shot learning
TSFSL	transductive supervised few-shot learning
ISSFSL	inductive semi-supervised few-shot learning
TSSFSL	transductive semi-supervised few-shot learning
$\mathcal{D}_{base}$, $\mathcal{D}_{novel}$	base data, novel data
$\mathcal{S}$, $\mathcal{Q}$, $\mathcal{U}$	support set, query set, unlabeled set
$\omega^r(\cdot)$, $\omega^m(\cdot)$	rotation-modal feature extractor, mirror-modal feature extractor
$\mathbf{A}$	adjacency matrix
$\mathbf{D}$	vertex degree matrix
$\mathbf{X}$	feature embedding of training data
$\mathbf{x}_{ts}$	feature embedding of testing data
$\mathbf{X}_s^r$, $\mathbf{X}_s^m$	support feature embedding on rotation-modal and mirror-modal
$\mathbf{X}_u^r$, $\mathbf{X}_u^m$	unlabeled feature embedding on rotation-modal and mirror-modal
$\mathbf{X}_q^r$, $\mathbf{X}_q^m$	query feature embedding on rotation-modal and mirror-modal
$\mathbf{x}_{select}^r$, $\mathbf{x}_{select}^m$	selected the most confidence samples' feature embedding on rotation-modal and mirror-modal
$\mathbf{Y}$	initial label matrix of training data
$\mathbf{Y}_s^r$, $\mathbf{Y}_s^m$	one-hot label matrices of support data on rotation-modal and mirror-modal
$\mathbf{Y}_u^r$, $\mathbf{Y}_u^m$	predicted soft-pseudo label matrices of unlabeled data on rotation-modal and mirror-modal
$\mathbf{Y}_q^r$, $\mathbf{Y}_q^m$	predicted soft label matrices of query data on rotation-modal and mirror-modal
$\mathbf{y}_{select}^r$, $\mathbf{y}_{select}^m$	selected the most confidence samples' one-hot-pseudo label vectors on rotation-modal and mirror-modal
$\mathbf{P}$	classifier
$\mathbf{P}^r$, $\mathbf{P}^m$	rotation-modal classifier, mirror-modal classifier
$\mathbf{P}_{opt}^r$, $\mathbf{P}_{opt}^m$	optimal rotation-modal classifier, optimal mirror-modal classifier

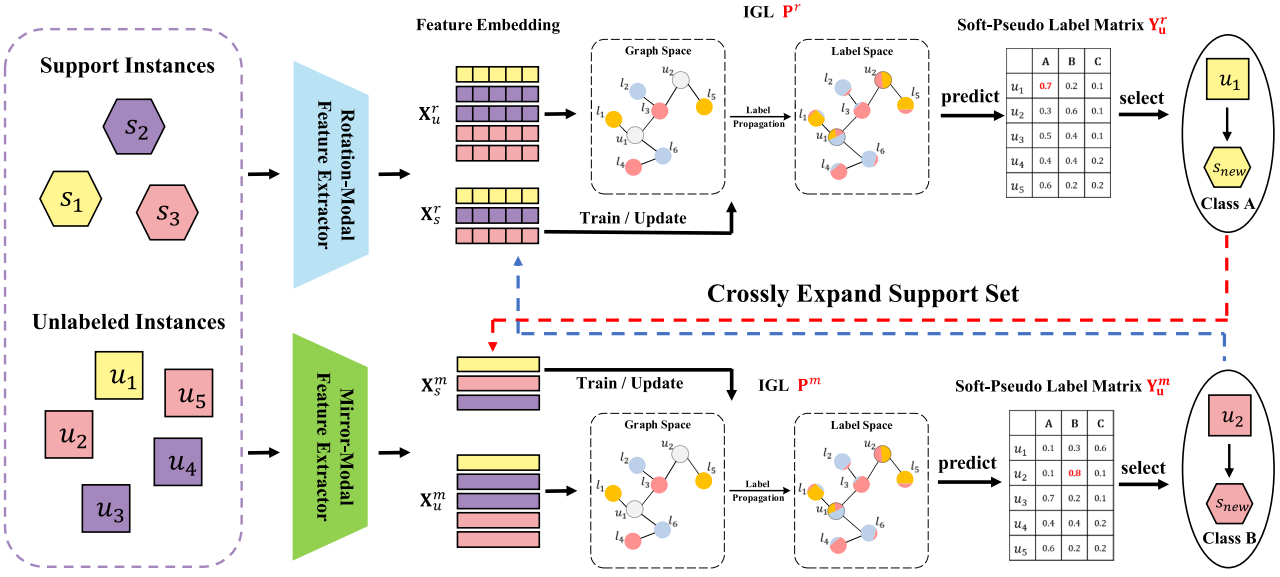


Fig. 2. The flowchart of graph co-training (GCT) in inductive semi-supervised case. We have two kinds of feature extractors, *i.e.*, rotation-modal feature extractor and mirror-modal feature extractor. $\mathbf{X}_s^r$ and $\mathbf{X}_s^m$ indicate the features of support data in rotation-modality and mirror-modality. $\mathbf{X}_u^r$ and $\mathbf{X}_u^m$ represent the features of unlabeled data in corresponding modalities. In each modality, we first employ the support samples to train the basic classifiers $\mathbf{P}^r$ and $\mathbf{P}^m$ (*i.e.*, IGL). Then, we use IGL classifier to predict unlabeled samples and obtain the corresponded soft-pseudo label matrices $\mathbf{Y}_u^r$ and $\mathbf{Y}_u^m$. Next, we select the most confident sample and give it one-hot-pseudo label. Note that different modalities may predict different results. Finally, we crossly expand the predicted-pseudo-sample to the support set and update the IGL classifier.

classifiers. To be more specific, their methods are based on the unified framework, which means that the feature extractors are usually guaranteed to match classifiers in a fixed way. Without this combination, the model performance will be significantly reduced, which greatly limits the scalability of the methods. While our GCT is freed with the feature extractor, thereby is more flexible to be applied in real scenarios.

C. Graph Learning

Graph Learning is an efficient way to model the data correlation of samples, composed of vertex set (*i.e.*, samples) and edge set. Each edge connects two vertices, which is capable of modeling pair-wise relations of samples. Researchers usually employ the adjacency matrix to represent a graph structure. [15] first proposed Graph Learning. In recent years, graph-based neural networks (GNN) have received extensive attention and have developed rapidly. Due to its good performance, GNN-based technologies have been applied in many fields, including few-shot learning [26]–[29], face clustering [30], [31], etc. The former achieves satisfactory performance by cooperating with meta-learning strategy, and the latter designs multiple kinds of GNNs to cooperate with each other to complete goals.

This paper focuses on traditional graph learning. In graph learning, researchers classify the unlabelled data by propagating the label information through graph structure, which is constructed by employing all the data (including labeled and unlabeled data). There are two challenges in the node/graph classification task by using graph learning: (i) In real applications, it is hard to know in advance what the to-be-tested sample looks like. It limits the scenario for this approach. (ii) Leading to a high computational cost. The predicting process is entirely online. All the data must be considered during the learning stage, which results in consuming massive computing resources. Furthermore, researchers have to re-construct the graph structure to propagate the label information when coming new testing data. This paper proposes a novel method dubbed as Isolated Graph Learning (IGL) to solve this challenge. IGL is a classifier which can be directly applied in the decoupled FSL task.

III. Problem Setup

In few-shot learning, there exist two processes (*i.e.*, pre-train process and meta-test process) with different categories of samples. Define the base data in pre-train phase as $\mathcal{D}_{base} = \{(x_{(i)}, y_{(i)})|y_{(i)} \in \mathcal{C}_{base}\}_{i=1}^{N_{base}}$, and the novel data in meta-test phase as $\mathcal{D}_{novel} = \{(x_{(j)}, y_{(j)})|y_j \in \mathcal{C}_{novel}\}_{j=1}^{N_{novel}}$, where x and y denote the sample and corresponded label. $\mathcal{C}_{base}$ and $\mathcal{C}_{novel}$ indicate the categories of base data and novel data, $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$. N_{base} and N_{novel} indicate the number of base data and novel data. To overcome overfitting, we follow the decoupled classification setups as [11]. There are three main stages in few-shot learning. First, we use $\mathcal{D}_{base}$ to train a CNN-based feature extractor $\omega(\cdot)$ in the pre-train phase. Second, we freeze the model's parameters and extract the feature embedding of $\mathcal{D}_{novel}$. Third, we design a classifier for the novel categories' classification.

Specifically, we first define the meta-test dataset as $\mathcal{D}_{novel} = \{\mathcal{S}, \mathcal{Q}, \mathcal{U}\}$, where $\mathcal{S}$, $\mathcal{Q}$, and $\mathcal{U}$ indicate support set, query set, and unlabeled set, $\mathcal{S} \cap \mathcal{Q} = \emptyset$, $\mathcal{S} \cap \mathcal{U} = \emptyset$, $\mathcal{Q} \cap \mathcal{U} = \emptyset$. Next, we divide the support, query, and unlabeled sets into different episodes. Each episode has $K\text{-}way\text{-}O\text{-}shot$ samples, where $K\text{-}way$ indicates K classes, and $O\text{-}shot$ denotes O samples per class. Finally, we employ the to-be-learned classifier to obtain the average classification accuracies of all the meta-test episodes on the query set $\mathcal{Q}$.

Besides, according to the differences of data used to construct the classifier, we split the few-shot learning into four kinds of settings: (i) inductive supervised few-shot learning (ISFSL), using the support features and labels to train the classifier; (ii) transductive supervised few-shot learning (TSFSL), using the support features, support labels, and query features to train the classifier; (iii) inductive semi-supervised few-shot learning (ISSFSL), using the support features, support labels, and unlabeled features to train the classifier; (iv) transductive semi-supervised few-shot learning (TSSFSL), using the support features, support labels, unlabeled features, and query features to train the classifier.

IV. Methodology

In this section, we describe our approach in detail. First, we encode the graph structure of samples' relations to the adjacency matrix. Second, we propose a novel graph learning method dubbed as Isolated Graph Learning (IGL) to tackle the FEM to some extent. This strategy introduces the Laplacian operator to transform the samples in feature space to graph representation and then project them to label space for prediction by regularization. Some details are shown in Figure 1. Third, we propose Graph Co-Training (GCT) by expanding the IGL to a co-training framework. GCT is capable of further addressing the FEM problem from a multi-modal fusion perspective. The flowchart is illustrated in Figure 2. At last, we introduce our designed feature extractors in detail.

A. Graph Structure Encoder

A graph can be formulated as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ represent vertex set and edge set, respectively. In the paper, the vertex set is composed of image samples. We encode the relations between edges and vertices through adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}$. Given the feature embedding of labeled vertices as $\mathbf{X} = [\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \cdots, \mathbf{x}_{(N)}] \in \mathbb{R}^{dim\times N}$, where $\mathbf{x}_{(i)}, (i = 1, 2, \cdots, N)$ indicates the embedding of v_i, and v_i denotes the i_{th} vertex in $\mathcal{V}$. dim and N denote the dimension and number of labeled samples. The elements in $\mathbf{A}$ can be defined as:

$$\mathbf{A}_{(i,j)} = \begin{cases} \exp\left(-dis\left(\mathbf{x}_{(i)}, \mathbf{x}_{(j)}\right)^2\right) & \text{if } (v_{(i)}, v_{(j)}) \in e \\ 0 & \text{o.w.} \end{cases} \quad (1)$$

where $(\cdot)_{(i,j)}$ is the (i,j)-element in $(\cdot)$. e indicates an edge in $\mathcal{E}$. $dis\left(\mathbf{x}_{(i)}, \mathbf{x}_{(j)}\right)$ represents the operator to calculate the distance of feature embeddings between $v_{(i)}$ and $v_{(j)}$, in our method, we select the k Nearest Neighbor (KNN) method. Following, we define the vertex degree matrix as $\mathbf{D} \in \mathbb{R}^{N\times N}$, which denotes a diagonal matrix with its (i, i)-element equal to the sum of the $i\text{-}th$ row of $\mathbf{A}$.

B. Isolated Graph Learning

In this section, we propose a novel label prediction method dubbed as Isolated Graph Learning (IGL). IGL is a strategy to solve the FEM problem by transforming the samples in

feature space to graph space. Unlike traditional graph learning, requiring both labeled and unlabelled data to construct the graph, the proposed IGL is more flexible to independently complete training and testing procedures by learning a regularized projection $\mathbf{P} \in \mathbb{R}^{dim \times C}$ to classify different categories. Here, C indicates the total number of classes. We calculate the cost function as:

$$\mathcal{F}(\mathbf{P}) = f_1(\mathbf{P}) + \lambda f_2(\mathbf{P}) + \mu f_3(\mathbf{P}) \tag{2}$$

where λ and μ represent the parameters to balance the function. $f_1(\mathbf{P})$ denotes the graph Laplacian regularizer, which can be formulated as:

$$\begin{aligned} f_1(\mathbf{P}) &= \frac{1}{2}\left(\sum_{i,j=1}^{N} \mathbf{A}_{(i,j)}\left(\frac{\left(\mathbf{X}^T\mathbf{P}\right)_{(i\cdot)}}{\sqrt{\mathbf{D}_{(i,i)}}} - \frac{\left(\mathbf{X}^T\mathbf{P}\right)_{(j\cdot)}}{\sqrt{\mathbf{D}_{(j,j)}}}\right)^2\right) \\ &= \mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\Delta\mathbf{X}^T\mathbf{P}\right) \end{aligned} \tag{3}$$

where $(\cdot)_{(i,\cdot)}$ is the $i\text{-}th$ row of $(\cdot)$. $\Delta = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ denotes the normalized graph Laplacian operator. $f_2(\mathbf{P})$ indicates the empirical loss term, which can be formulated as:

$$f_2(\mathbf{P}) = \left\|\mathbf{X}^T\mathbf{P} - \mathbf{Y}\right\|_F^2 \tag{4}$$

where $\mathbf{Y} \in \mathbb{R}^{N \times C}$ indicates the initial label embedding matrix. For labeled samples, if the $i\text{-}th$ sample belongs to the $j\text{-}th$ class, $\mathbf{Y}_{(i,j)}$ is 1, and otherwise, it is 0. $f_3(\mathbf{P})$ is the constraint term. In this paper, we introduce $\ell_{2,1}\text{-}norm$ to select an essential feature and avoid overfitting for $\mathbf{P}$, which can be defined as:

$$f_3(\mathbf{P}) = \|\mathbf{P}\|_{\ell_{2,1}} \tag{5}$$

where $\|\cdot\|_{\ell_{2,1}}$ represents $\ell_{2,1}$-norm of $(\cdot)$. The objective function on IGL can be formulated as:

$$\begin{aligned} &\arg\min_{\mathbf{P}} \mathcal{F}(\mathbf{P}) \\ &\quad = \mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\Delta\mathbf{X}^T\mathbf{P}\right) + \lambda\left\|\mathbf{X}^T\mathbf{P} - \mathbf{Y}\right\|_F^2 + \mu\|\mathbf{P}\|_{\ell_{2,1}} \end{aligned} \tag{6}$$

To optimize this problem, we first relax Equation 6 as:

$$\begin{aligned} &\arg\min_{\mathbf{P},\mathbf{B}} \mathcal{F}(\mathbf{P},\mathbf{B}) \\ &\quad = \mathrm{tr}\left(\mathbf{P}^T\mathbf{X}\Delta\mathbf{X}^T\mathbf{P}\right) + \lambda\left\|\mathbf{X}^T\mathbf{P} - \mathbf{Y}\right\|_F^2 + \mu\,\mathrm{tr}\left(\mathbf{P}^T\mathbf{B}\mathbf{P}\right) \end{aligned} \tag{7}$$

where $\mathbf{B}$ is a diagonal matrix. Then we alternately update $\mathbf{P}$ and $\mathbf{B}$ until Equation 7 convergence, follow [32], we directly solve the problems as:

$$\mathbf{B}_{(i,i)} = \frac{1}{2\left\|\mathbf{P}_{(i,\cdot)}\right\|_2^2 + 10^{-8}},\ i = 1, \ldots, dim \tag{8}$$

$$\mathbf{P} = \mathcal{H}(\mathbf{X}) = \lambda\left(\mathbf{X}\Delta\mathbf{X}^T + \lambda\mathbf{X}\mathbf{X}^T + \mu\mathbf{B}\right)^{-1}\mathbf{X}\mathbf{Y} \tag{9}$$

Following, given a testing sample embedding $\mathbf{x}_{ts} \in \mathbb{R}^{dim \times 1}$, we predict the $\mathbf{x}_{ts}$'s category by:

$$\mathcal{Z}(\mathbf{x}_{ts}) = id_{max}\left\{\mathbf{x}_{ts}{}^T\mathbf{P}\right\} \tag{10}$$

where id_{max} represents an operator to obtain the index of the max value in the vector.

Algorithm 1: Graph Co-Training for ISSFSL

Input: Base data $\mathcal{D}_{base}$, novel data $\mathcal{D}_{novel}$
Output: Query label
 # pre-train phase
Employ $\mathcal{D}_{base}$ to train the rotation-modal feature extractor $\omega^r(\cdot)$ and mirror-modal feature extractor $\omega^m(\cdot)$.
 # Meta-test phase
Extract two-modal feature of $\mathcal{D}_{novel}$ by $\mathbf{X}^r_{novel} = \omega^r(\mathcal{D}_{novel})$, and $\mathbf{X}^m_{novel} = \omega^m(\mathcal{D}_{novel})$.
 # Co-training steps:
Intialize $\mathbf{B}$ and $\mathbf{P}$.
repeat
  Construct two modal classifiers by Equation 11.
  Infer pseudo labels and enlarge support data by Equation 12, 13.
until *the unlabeled data is exhausted.*
Predict the query labels by Equation 14.

C. Graph Co-Training for Few-Shot Learning

As mentioned in Section III, there exist four kinds of setting in FSL. To make our IGL perform well in all kinds of FSL, we introduce the co-training strategy to further cooperate with IGL, and named the new approach as Graph Co-Training (GCT). On the one hand, it can solve the FEM problem from a multi-modal fusion perspective. For the details of multi-modal information, please refer to Section IV-E. On the other hand, GCT is capable of strengthening the robustness of the to-be-learned classifier by employing the unlabeled data $\mathcal{U}$. We'll detail how GCT works in the four settings. Notably, both the construction of $\mathbf{P}$ and the collaborative training are designed for each episode.

First see the inductive semi-supervised few-shot learning (ISSFSL). We construct the co-training framework with two modal features, *i.e.*, rotation-modality and mirror-modality. The rotation-model feature extractor, $\omega^r(\cdot)$ follows [16], the corresponding embeddings of novel data can be defined as $\mathbf{X}^r_{novel} = \omega^r(\mathcal{D}_{novel}) = [\mathbf{X}^r_s, \mathbf{X}^r_u, \mathbf{X}^r_q]$, where $\mathbf{X}^r_s = \omega^r(\mathcal{S})$, $\mathbf{X}^r_u = \omega^r(\mathcal{U})$, and $\mathbf{X}^r_q = \omega^r(\mathcal{Q})$ indicate the features of support, unlabeled, and query data on the rotation-model. The mirror-modal feature extractor, $\omega^m(\cdot)$ follows [17], the features in this modal are denoted as $\mathbf{X}^m_{novel} = \omega^m(\mathcal{D}_{novel}) = [\mathbf{X}^m_s, \mathbf{X}^m_u, \mathbf{X}^m_q]$. The complete GCT is demonstrated in Figure 2, which consists of four steps:

(i) From Equation 9, we construct two different classifiers $\mathbf{P}^r$ and $\mathbf{P}^m$ by employing two modal support features $\mathbf{X}^r_s$ and $\mathbf{X}^m_s$, respectively.

$$\begin{cases} \mathbf{P}^r = \mathcal{H}(\mathbf{X}^r_s) \\ \mathbf{P}^m = \mathcal{H}(\mathbf{X}^m_s) \end{cases} \tag{11}$$

(ii) Predict the unlabeled data's label from two modal features by:

$$\begin{cases} \mathbf{Y}^r_u = \mathbf{X}^{r\,T}_u\mathbf{P}^r \\ \mathbf{Y}^m_u = \mathbf{X}^{m\,T}_u\mathbf{P}^m \end{cases} \tag{12}$$

where $\mathbf{Y}^r_u$ and $\mathbf{Y}^m_u$ denote predicted **soft-pseudo label matrices** of unlabeled data on rotation-modal and mirror-modal.

(iii) Rank the values in soft-pseudo label matrices, then selecting the most confident unlabeled samples' feature $\mathbf{x}^r_{select}$ and $\mathbf{x}^m_{select}$ on each modal, then asserting them corresponding **one-hot-pseudo label vectors** $\mathbf{y}^r_{select}$ and $\mathbf{y}^m_{select}$ (for more details about how to select the most confident sample, please refer to Section IV-D). Next, crossly extend the pseudo-labeled samples and corresponding labels to the support set on different modals. We formulate this step as:

$$\begin{cases} \mathbf{X}^r_s = [\mathbf{X}^r_s, \mathbf{x}^m_{select}], \mathbf{Y}^r_s = [\mathbf{Y}^r_s, \mathbf{y}^m_{select}] \\ \mathbf{X}^m_s = [\mathbf{X}^m_s, \mathbf{x}^r_{select}], \mathbf{Y}^m_s = [\mathbf{Y}^m_s, \mathbf{y}^r_{select}] \end{cases} \tag{13}$$

where $\mathbf{Y}^r_s$ and $\mathbf{Y}^m_s$ denote the **one-hot label matrices** of support data on two models.

(iv) Repeat (i), (ii) (iii) until the unlabeled data is exhausted (in the real application, we usually select 80 unlabeled samples, for more discussions and results, please see Section V-C.2 and Figure 7). Then we obtain two optimal classifiers $\mathbf{P}^r_{opt}$ and $\mathbf{P}^m_{opt}$. Employ them to predict the query labels by:

$$\mathcal{Z}(\mathbf{X}^r_q, \mathbf{X}^m_q) = id_{max}\left\{\frac{\left(\mathbf{X}^{r\,T}_q \mathbf{P}^r_{opt} + \mathbf{X}^{m\,T}_q \mathbf{P}^m_{opt}\right)}{2}\right\} \tag{14}$$

where $\mathbf{Y}^r_q = \mathbf{X}^{r\,T}_q \mathbf{P}^r_{opt}$ and $\mathbf{Y}^m_q = \mathbf{X}^{m\,T}_q \mathbf{P}^m_{opt}$, denote the predicted **soft label matrices** of query data on rotation-modality and mirror-modality. We summarize the steps in Algorithm 1.

Next see the transductive semi-supervised few-shot learning (TSSFSL). The query feature is also available when constructing the classifier. Here, we can implement TSSFSL with only a few minor tweaks. Specifically, in step (ii), we have to predict not only unlabeled data but also query data.

Then see the transductive supervised few-shot learning (TSFSL), the unlabeled data is not available, but the query feature is given in advance. Therefore, we just need to replace the unlabeled data with query data in (i)(ii)(iii) steps, and finally classify the query data.

At last see the inductive supervised few-shot learning (ISFSL), the unlabeled data is unavailable, and query data is not given to us in advance. Thus, the co-training strategy cannot be used in this case and we think the basic IGL classifier is optimal. We can complete ISFSL by Equation 11 and 14.

D. How to Select the Most Confident Sample?

Here we take the rotation-modal as an example to illustrate the strategy. According to Equation 12, we can get the predicted rotation-model soft label matrix $\mathbf{Y}^r_u \in \mathbb{R}^{N_u \times C}$, where N_u denotes the number of unlabeled samples; C denotes the number of unlabeled categories. For each element $\mathbf{Y}^r_{u(n,c)}$, it means the probability of $n\text{-}th$ sample belongs to the $c\text{-}th$ category, where $n = 1, 2, \cdots, N_u$, $c = 1, 2, \cdots, C$. In our strategy, we first traverse all the elements in $\mathbf{Y}^r_u$ to find the largest element, which can be defined as $\mathbf{Y}^r_{u(n_{max}, c_{max})}$. The $n_{max}\text{-}th$ sample is the to-be-selected most confident sample, which belongs to the $c_{max}\text{-}th$ category.

E. Multi-Modal Feature Extractor

We can adopt various modal features from different feature extractors to achieve our purpose. For example: (i) Standard modality (Std-Mod) [11], the feature extractor comes from a standard CNN-based classification structure. (ii) Meta modality (Meta-Mod) [33], the feature extractor combines the strategy of meta-learning with the network. (iii) Self-supervised rotation modality (SS-R-Mod) [16], the feature extractor introduces auxiliary loss to predict the angle of image rotation, including $\{0°, 90°, 180°, 270°\}$. (iv) Self-supervised mirror modality (SS-M-Mod) [17]. Different from the SS-R-Mod, SS-M-Mod introduces another auxiliary loss to predict image mirrors, including $\{vertically, horizontally, diagonally\}$. In most of the experiments, we present the results of collaborative training with SS-R-Mod and SS-M-Mod, and thereby we briefly introduce them.

In the SS-R-Mod, the feature extractor updates the network parameters with two kinds of loss, which are the standard classification loss (*i.e.*, $\mathcal{L}^s$) and rotation-based self-supervised auxiliary loss (*i.e.*, $\mathcal{L}^r$). To be more specific, assume there's a base image feature vector $\mathbf{x}$. We project it into a label space, *i.e.*, $\mathbf{x} \to \mathbf{z}^s$, where $\mathbf{z}^s = [z^s_1, z^s_2, \cdots, z^s_{C_{base}}] \in \mathbb{R}^{C_{base}}$, C_{base} denotes the number of the base category. Then transform it to the probability distribution and calculate the standard classification loss $\mathcal{L}^s$ by cross-entropy function:

$$\mathcal{L}^s = -\sum_{c=1}^{C_{base}} \hat{y}^s_c log(y^s_c) \tag{15}$$

where $y^s_c = \frac{e^{z^s_c}}{\sum_{c=1}^{C_{base}} e^{z^s_c}}$ indicates the predicted probability that sample $\mathbf{x}$ belongs to class c, while $\hat{y}^s_c$ denotes the groundtruth probability. After that, we map the $\mathbf{x}$ to rotation-based label space, *i.e.*, $\mathbf{x} \to \mathbf{z}^r$, where $\mathbf{z}^r = [z^r_1, z^r_2, z^r_3, z^r_4] \in \mathbb{R}^4$. We can get the auxiliary loss by:

$$\mathcal{L}^r = -\sum_{c=1}^{4} \hat{y}^r_c log(y^r_c) \tag{16}$$

where $y^r_c = \frac{e^{z^r_c}}{\sum_{c=1}^{4} e^{z^r_c}}$ indicates the predicted probability of rotation angle, while $\hat{y}^r_c$ denotes the groundtruth probability. The complete loss in SS-R-Mod is $\mathcal{L}^s + \mathcal{L}^r$.

In the SS-M-Mod, it also uses the standard classification loss, but change the rotation-based self-supervised auxiliary loss to mirror-based self-supervised auxiliary loss $\mathcal{L}^m$, which can be defined as:

$$\mathcal{L}^m = -\sum_{c=1}^{3} \hat{y}^m_c log(y^m_c) \tag{17}$$

where y^m_c indicates the predicted probability of mirroring way, while $\hat{y}^m_c$ denotes the groundtruth probability. The complete loss in SS-M-Mod is $\mathcal{L}^s + \mathcal{L}^m$.

F. Discussion and Analysis

In this section, we will further discuss and analyze our work from two aspects. We first see the comparison of our Isolated Graph Learning (IGL) with traditional Graph Learning (GL); next conclude with a comprehensive analysis of the reasons for the success of GCT.

1) Comparison of IGL and GL: Compared with IGL, the GL's objective function is different, which can be formulated as:

$$\mathcal{F}(\mathbf{R}) = f_4(\mathbf{R}) + \beta f_5(\mathbf{R}) \tag{18}$$

where $\mathbf{R}$ is the to-be-predicted soft label matrix. β denotes the parameter to balance the function. $f_4(\mathbf{R})$ denotes the graph Laplacian regularizer, which can be formulated as:

$$f_4(\mathbf{R}) = \mathrm{tr}\left(\mathbf{R}^T \Delta \mathbf{R}\right) \tag{19}$$

where $\Delta = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ denotes the normalized graph Laplacian operator. $f_5(\mathbf{R})$ indicates the empirical loss term, which can be formulated as:

$$f_5(\mathbf{R}) = \|\mathbf{R} - \mathbf{Y}\|_F^2 \tag{20}$$

where $\mathbf{Y}$ indicates the initial one-hot label matrix.

Comparing equations 2,3,4 and equations 18,19,20, it can be seen that the most essential difference between IGL and GL is that we replace $\mathbf{R}$ with $\mathbf{X}^T\mathbf{P}$. If directly using $\mathbf{R}$, the training samples and the to-be-classified testing samples must be employed together to achieve the testing samples' label prediction. In other words, if we only use the training data to construct the graph, we can only achieve label propagation between training samples. When a new batch of testing data arrives, we must rebuild the graph based on the training and testing data to realize the label prediction for the testing samples. While IGL is different. After we get $\mathbf{P}$ based on the training data, we can directly predict the label of new testing data based on $\mathbf{P}$ without rebuilding the graph.

2) Comprehensive Analysis of GCT: Here, let's sort out why GCT is effective. After dismantling it, we find that there are three parts that positively influence our method:

(i) The first point is that the base classifier IGL we designed maps the original features to the graph space, which can reduce the dependence on features in the FSL task, thereby weakening the influence of the FEM problem.

(ii) The second point is that we introduce multi-modal information. In the FSL task, the features obtained by different feature extractors are different. Although they all cause the distribution-shift, the angle of shift varies. Through the mutual correction of multi-modal information, the final performance can be improved.

(iii) The third point is that the employed co-training strategy is reasonable and efficient. As mentioned before, we use multi-modal information here. However, multi-modal features are a double-edged sword. If used well, features with different shortcomings can be corrected with each other to improve performance. If used incorrectly, the performance will be further degraded. In the co-training strategy, we select the most confident sample in a single modality (the selected sample can be treated as the one that is not affected by distribution-shift), and then amplify the advantages of each modality through the alternate iterations of the two modalities, so as to enhance the classifier's ability.

V. Experiments

In this section, we design experiments to evaluate our method. Specifically, we first illustrate the experimental setup, containing datasets and implementation details. Then, we demonstrate the comparison results and discuss them in detail. Next, we design ablation studies to further analyze our method. In the end, we observe the experimental performances of multi-modal fusion and cross-domain. All experiments are conducted on a Tesla-V100 GPU with $32G$ memory.

A. Experimental Setup

1) Datasets: Our experiments are carried out on five benchmark datasets, including mini-ImageNet [56], tiered-ImageNet [34], CIFAR-FS [33], FC100 [42], and CUB [57]. mini-ImageNet and tiered-ImageNet are selected from the ImageNet dataset [58] and as the subsets. mini-ImageNet consists of 100 classes with 600 images per class, and tiered-ImageNet has 608 classes and each class contains $1,281$ images on average. Both of them resize the image to 84×84. Following the standard split way as [11], for mini-ImageNet, the base set contains 64 selected classes, the validation is composed of 16 classes, and the novel set includes 20 classes. Similarly, for tiered-ImageNet, the base set includes 351 classes, the validation set contains 97 classes, and 160 classes are prepared for the novel set. The CIFAR-FS and FC100 are the subsets of the CIFAR-100 dataset [59], which includes 100 classes. According to the split introduced in [33], CIFAR-FS is divided into 64 classes, and it can be seen as the base set, the validation set consists of 16 classes, and the novel set includes 20 classes. And for FC100, we divided it into 60 classes as the base set, the validation set contains 20 classes, and the novel set includes 20 classes. The image size of CIFAR-FS and FC100 datasets are set to 32×32. There are $11,788$ images with 200 categories in the CUB dataset in total. Referring to the implementation in ICI [11], CUB is divided into 100 classes as the base set, the validation set contains 50 classes, and the novel set consists of 50 classes. In all experiments, the images are cropped into 84×84.

2) Implementation Details: In our paper, both the rotation-modal and mirror-modal feature extractors adopt the ResNet12 [60] backbone, which contains four residual blocks (a convolution layer with 3×3 kernel size, batch normalization layer, and LeakyReLU layer), four 2×2 max-pooling layers, and four dropout layers. The optimizer is adopting stochastic gradient descent (SGD) with Nesterov momentum (0.9). The training epochs are set as 120, and then all module was tested over 600 episodes with 15 query samples per class. For more details of network experimental implements, such as the learning rate, data augmentation, and the number of filters, please refer to ICI [11]. Besides the feature extractor, the designed classifier also affects the final results (*i.e.*, the two parameters in Equation 6). For fairness and convenience, we fix $\lambda = 0.1$ and $\mu = 0.6$ for all the datasets by our empirical tuning. And in Section V-C.1, we discuss the influence of the two parameters.

TABLE II

THE 5-WAY SEMI-SUPERVISED FEW-SHOT CLASSIFICATION ACCURACIES (%) WITH 95% CONFIDENCE INTERVALS OVER 600 EPISODES. $(\cdot)^{\star}$, AND $(\cdot)^{\dagger}$ IN THIS TABLE INDICATE INDUCTIVE SEMI-SUPERVISED, AND TRANSDUCTIVE SEMI-SUPERVISED SETTINGS, RESPECTIVELY. ABOUT THE INFLUENCE OF THE UNLABELED NUMBER, PLEASE SEE FIGURE 7

Method	Venue	Backbone	mini-ImageNet		tiered-ImageNet		CIFAR-FS		FC100	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MSkM [34]	ICLR'18	ResNet12	62.10	73.60	68.60	81.00	-	-	-	-
TPN [35]	ICLR'19	ResNet12	62.70	74.20	72.10	83.30	-	-	-	-
LST [10]	NIPS'19	ResNet12	70.10	78.70	77.70	85.20	73.00	85.62	42.77	57.67
EPNet [36]	ECCV'20	ResNet12	75.36	84.07	81.79	88.45	76.77	86.03	45.21	59.81
TransMatch [37]	CVPR'20	ResNet12	63.02	81.19	-	-	-	-	-	-
ICI [11]	CVPR'20	ResNet12	71.41	81.12	85.44	89.12	78.07	84.76	46.27	61.30
PLCM [12]	ICCV'21	ResNet12	72.06	83.71	84.78	90.11	77.62	86.13	48.35	62.75
PTN [38]	AAAI'21	WRN	**82.66**	88.43	84.70	89.14	-	-	-	-
iLPC [13]	ICCV'21	ResNet12	70.99	81.06	85.04	89.63	78.57	85.84	-	-
MHFC [17]	ACMMM'21	ResNet12	79.26	87.30	87.57	91.80	84.74	90.19	50.95	64.05
GCT*	Ours	ResNet12	78.73	**88.57**	88.65	91.53	84.91	90.67	52.78	**65.21**
GCT†	Ours	ResNet12	80.04	88.29	**88.74**	**92.07**	**85.11**	**91.00**	**53.64**	64.95

TABLE III

THE 5-WAY MULTI-MODAL FUSION BASED FEW-SHOT CLASSIFICATION ACCURACIES (%) WITH 95% CONFIDENCE INTERVALS OVER 600 EPISODES IN MINI-IMAGENET AND TIERED-IMAGENET. FOR A FAIR COMPARISON, THE REPORTED RESULTS OF OUR GCT IS BASED ON THE TRANSDUCTIVE SUPERVISED SETTING

Method	Backbone	mini-ImageNet		tiered-ImageNet	
		1-shot	5-shot	1-shot	5-shot
DenseCls [22]	ResNet12	62.53	79.77	-	-
DWC [25]	ResNet12	63.73	81.19	70.44	85.43
DivCoop [23]	ResNet12	64.14	81.23	-	-
URT [24]	ResNet12	72.23	83.35	80.30	88.63
MHFC [17]	ResNet12	73.10	81.75	82.10	87.99
GCT	ResNet12	**75.29**	**85.17**	**86.10**	**90.18**

TABLE IV

THE 5-WAY FEW-SHOT CLASSIFICATION ACCURACIES (%) WITH 95% CONFIDENCE INTERVALS OVER 600 EPISODES IN CIFAR-FS AND FC100. THIS TABLE COMPARES OUR GCT WITH STATE-OF-THE-ARTS WITHOUT CONSIDERING ANY VARIABLES, SUCH AS BACKBONES, TRICKS, OR EVEN THE FEW-SHOT SETTINGS, JUST REPORTS THE FINAL PERFORMANCES

Method	Backbone	CIFAR-FS		FC100	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [39]	4CONV	55.50	72.00	35.30	48.600
MAML [40]	4CONV	58.90	71.50	-	-
RelationNet [41]	4CONV	55.00	69.30	-	-
TADAM [42]	ResNet12	-	-	40.10	56.10
DenseCls [22]	ResNet12	-	-	42.04	57.63
MetaOpt [33]	ResNet12	72.00	84.20	41.10	55.50
TEAM [43]	ResNet12	70.43	81.25	-	-
MABAS [44]	ResNet12	73.24	85.65	41.74	57.11
Fine-tuning [45]	WRN	76.58	85.79	43.16	57.57
DSN-MR [46]	ResNet12	75.60	86.20	-	-
GCT	ResNet12	**85.11**	**91.00**	**53.64**	**65.21**

B. Experimental Results

1) Comparison Results With Semi-Supervised Methods: Table II shows the comparison results with recently proposed semi-supervised few-shot classification methods. These approaches use the unlabeled samples to correct the distribution. Obviously, our GCT achieves outstanding performances. First see the comparison without considering PTN [38]. Specifically, in mini-ImageNet, GCT exceeds others 0.78%-17.94% in the 1-shot case, and 1.27%-14.97% in the 5-shot case; in tiered-ImageNet, GCT exceeds others 1.17%-20.14% in the 1-shot case, and 0.27%-11.07% in the 5-shot case; in CIFAR-FS, GCT exceeds others 0.37%-12.11% in the 1-shot case, and 0.81%-6.24% in the 5-shot case; in FC100, GCT exceeds others 2.69%-10.87% in the 1-shot case, 1.16%-7.54% in the 5-shot case. And then comparing GCT with PTN, we find that only in mini-ImageNet, the performance of GCT is slightly inferior to PTN, while in other datasets, GCT is better than PTN.

2) Comparison Results With Multi-Modal Fusion Methods: Our GCT fuses two modal information, so that it is necessary to compare it with recently proposed multi-modal fusion based methods. The comparison results are presented in Table III. For a fair comparison, we do not utilize the unlabeled data. We exploit our GCT in the transductive setting. We find that our GCT can outperform other methods by 1.82%-12.76% in mini-ImageNet, and 1.55%-15.66% in tiered-ImageNet.

3) Comparison Results With State-of-The-Art Methods: All the works have their highlights and tricks to improve their final performances. In this section, we compare our GCT with other state-of-the-art methods. That means, we do not consider the impact of any variables, such as backbones, various tricks, or even the few-shot settings (*i.e.*, inductive supervised, transductive supervised, inductive semi-supervised, transductive semi-supervised settings), but only report their final results. From Table V,IV, we observe that, our proposed GCT has achieved significant improvements compared with other SOTAs. In mini-ImageNet dataset, GCT exceeds others at least 11.47% in the 1-shot case, and 5.17% in the 5-shot case. In tiered-ImageNet dataset, GCT outperforms others at least 7.95% in the 1-shot case, and 2.17% in the 5-shot case. In CIFAR-FS dataset, GCT gains improvements of at least 8.53% in the 1-shot case, and 4.52% in the 5-shot case. In FC100 dataset, GCT surpasses others by at least 10.36% in the 1-shot case, and 6.88% in the 5-shot case.

C. Ablation Studies

In the paper, we make two efforts to address the FEM problem. First, we propose the IGL classifier to encode

TABLE V

THE 5-WAY FEW-SHOT CLASSIFICATION ACCURACIES (%) WITH 95% CONFIDENCE INTERVALS OVER 600 EPISODES IN MINI-IMAGENET AND TIERED-IMAGENET. THIS TABLE COMPARES OUR GCT WITH SEVERAL STATE-OF-THE-ARTS WITHOUT CONSIDERING ANY VARIABLES, SUCH AS BACKBONES, TRICKS, OR EVEN THE FEW-SHOT SETTINGS, JUST REPORTS THE FINAL PERFORMANCES

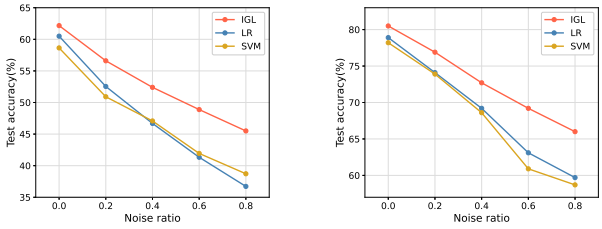
Method	Backbone	mini-ImageNet		tiered-ImageNet	
		1-shot	5-shot	1-shot	5-shot
ProtoNet [39]	4CONV	49.42	68.20	-	-
MAML [40]	4CONV	48.70	63.11	-	-
RelationNet [41]	ResNet18	52.48	69.83	-	-
Baseline [47]	ResNet18	51.75	74.27	-	-
Baseline++ [47]	ResNet18	51.87	75.68	-	-
LEO [48]	WRN	61.76	77.59	66.33	81.44
TPN [35]	4CONV	52.78	66.42	55.74	71.01
AM3 [49]	ResNet12	65.30	78.10	69.08	82.58
TapNet [50]	ResNet12	61.65	76.36	63.08	80.26
CTM [51]	ResNet18	64.12	80.51	-	-
MetaOpt [33]	ResNet12	62.64	78.63	65.99	81.56
TEAM [43]	ResNet12	60.07	75.90	-	-
S2M2 [16]	WRN	64.93	83.18	73.71	88.59
Fine-tuning [45]	WRN	65.73	78.40	73.34	85.50
DSN-MR [46]	ResNet12	64.60	79.51	67.39	82.85
MABAS [44]	ResNet12	64.21	81.01	-	-
HGNN [52]	4CONV	60.03	79.64	64.32	83.34
DC [53]	WRN	68.57	82.88	78.19	89.90
ICI [11]	ResNet12	66.80	79.26	80.79	87.92
MELR [54]	ResNet12	67.40	83.40	72.14	87.01
ODE [55] 1	ResNet12	67.76	82.71	71.89	85.96
GCT	ResNet12	**80.04**	**88.57**	**88.74**	**92.07**

the sample's feature embedding to the graph representation. Second, we propose GCT block, which extends IGL to the semi-supervised setting and fuses two-modal information. To this end, we design ablation studies to analyze how the two blocks affect the results. Besides, there are two important tricks in our method, *i.e.*, feature-alignment and fine-tuning. We also evaluate their efficiency to our method in this section.

1) Influence of IGL Classifier: (i) In this paper, we have pointed out the reason to design the graph-based classifier: Transforming the raw data to graph representation is helpful to reduce the dependence on features in classification tasks, thereby suffering less from feature noise. To evaluate this conclusion, we deliberately introduce different degrees of Gaussian noise to the features, and compare our IGL with the graph-irrelevant classifiers, including Logistic Regression (LR) and Support Vector Machine (SVM). The results with inductive supervised setting are presented in Figure 3. Obviously, our IGL is more insensitive to noise, and achieves better performances on both 5-way 1-shot and 5-way 5-shot cases.

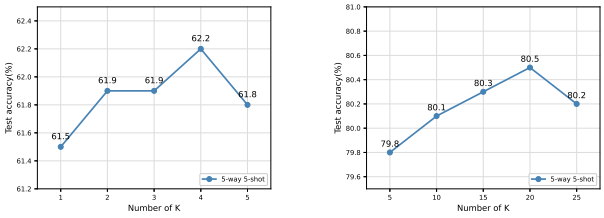
To better understand the phenomenon, we give an explanation here: In the LR and SVM, the features are directly used to construct the classifiers; while in the GCT, the features are first be used to construct the adjacency matrix $\mathbf{A}$ through Equation 1, and then the $\mathbf{A}$ is used to construct the classifier by cooperating with graph Laplacian operator through Equation 3, 9. That means, the graph-based classifiers rely not only on the quality of the features, but also on the way of constructing $\mathbf{A}$.

(ii) Building a high-quality graph structure is the cornerstone of the success of our approach. In this paper, we adopt



(a) 5-way 1-shot (b) 5-way 5-shot

Fig. 3. Comparison results with different noise ratios. The experiments are conducted in mini-ImageNet with inductive supervised setting. These experiments do not use the feature-alignment and fine-tuning tricks.



(a) 5-way 1-shot (b) 5-way 5-shot

Fig. 4. Comparison results of different K in KNN-based graph-constructing method. The experiments are conducted in mini-ImageNet with inductive supervised setting. These experiments do not use the feature-alignment and fine-tuning tricks.

TABLE VI

COMPARISON RESULTS OF EMPLOYING DIFFERENT GRAPH-CONSTRUCTING METHODS. THE EXPERIMENTS ARE CONDUCTED IN MINI-IMAGENET WITH INDUCTIVE SUPERVISED SETTING. THESE EXPERIMENTS DO NOT USE THE FEATURE-ALIGNMENT AND FINE-TUNING TRICK

Method	mini-ImageNet	
	5-way 1-shot	5-way 5-shot
IGL with KNN	62.17	80.52
IGL with KMeans	61.73	79.65

the KNN method to construct the graph. It is crucial to choose a reasonable K. Here, we show the influence of K in Figure 4. The experiments are conducted in mini-ImageNet with inductive supervised setting. We observe that K=3 in the 1-shot case and K=20 in the 5-shot case are the optimal selection. In practice, we will first simulate the real scene with the base data, select the appropriate K and apply it directly to the corresponding task.

In order to further evaluate the influence of the graph-constructing method on the results, we use another classical method KMeans to complete related tasks. In the KNN-based method, each vertex will have the same number of edges, that is, the degrees of the vertices are the same, while the KMeans-based method does not have this constraint. The comparison results are listed in Table VI. We find that the KMeans-based method is slightly inferior in the FSL task.

(iii) From Equation 6, we know that there exist two parameters (*i.e.*, λ and μ) influence the performance. Let's fix one to see how the other changes. The results of four datasets are listed in Figure 6. All the experiments are conducted in the inductive supervised setting. From the results, we find our IGL is not sensitive to the parameters.

2) Influence of GCT Block: (i) The IGL can be extended to semi-supervised tasks through the proposed GCT block. The

TABLE VII

ABLATION STUDY TO EVALUATE THE INFLUENCE OF DIFFERENT TRICKS (*feature-alignment* AND *fine-tuning*) IN INDUCTIVE SEMI-SUPERVISED CASE WITH 80 UNLABELED SAMPLES

GCT	Feature-alignment	Fine-tuning	mini-ImageNet	
			5-way 1-shot	5-way 5-shot
✓			75.93	86.43
✓	✓		78.44	87.95
✓		✓	77.42	87.63
✓	✓	✓	78.73	88.57

TABLE VIII

COMPARISON RESULTS OF DIFFERENT MODALS IN MINI-IMAGENET. WHEN ADOPTING SELF-TRAINING OR CO-TRAINING STRATEGIES, THE USED UNLABELED SAMPLES ARE 15. IGL IS THE CLASSIFIER. ST-R DENOTES USING THE FEATURE OF SS-R-MOD; ST-M DENOTES USING THE FEATURE OF SS-M-MOD; ST-R-M DENOTES FUSING ROTATION-BASED AND MIRROR-BASED SELF-SUPERVISED STRATEGIES TO ONE BRANCH IN THE PRE-TRAINING STAGE. ④ DENOTES DIRECTLY FUSING THEIR DECISIONS THROUGH EQUATION 14. THESE EXPERIMENTS DO NOT USE THE FEATURE-ALIGNMENT AND FINE-TUNING TRICKS

Method	mini-ImageNet	
	5-way 1-shot	5-way 5-shot
① ST-R + IGL	61.23	78.12
② ST-M + IGL	60.94	78.65
③ ST-R-M + IGL	61.77	79.98
④ ST-R + ST-M + IGL	62.17	80.52
⑤ ST-R + IGL + Self-Training	71.01	82.50
⑥ ST-M + IGL + Self-Training	71.15	82.64
⑦ ST-R-M + IGL + Self-Training	71.90	83.07
⑧ ST-R + ST-M + IGL + Co-Training	73.12	83.63

employed number of unlabeled samples plays a critical role in our final performance. We report the comparison results in Figure 7, the x-axis represents the number of unlabeled samples. Obviously, with the increasing of unlabeled samples, the performance of our method has become more competitive. Meanwhile, the performance will reach saturation or even decline after 50 unlabeled samples.

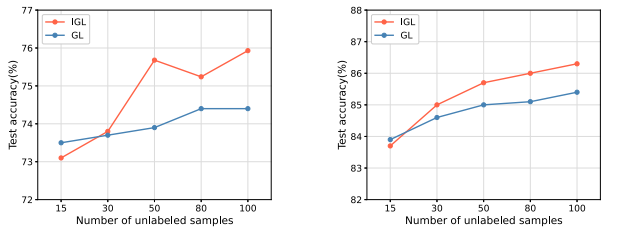
(ii) Besides, GCT is a powerful method to fuse two modal information. Therefore, we design ablation studies to observe the single-modal performances. We demonstrate the results in Table VIII. Experimental results demonstrate the effectiveness of the fusion strategy to a certain extent. To estimate the influence of *fusion* further, we demonstrate the results of each class. To be more specific, we select one episode (including 5 classes) on a 5-shot case, randomly. And then visualize the corresponding confusion matrices of each modal in Figure 8,9. Obviously, different results are obtained from various modal features for a specific class. And the proposed GCT achieves competitive results, reaching the same as the best performance result of a single model. Therefore, with the increase of category number, the performance of the proposed method also improves and is naturally more favorable.

3) Influence of Tricks: Each classical paper has the corresponding tricks to improve the final performance, and so do we. (i) Since the to-be-fused features come from different separate spaces, it consists of inconsistent measurement scales problem. Therefore, before starting our GCT operation, we attempt to align the features through the conventional subspace learning algorithm to transfer the initial features to a unified space with reconstructed low-dimensional representation. Specifically, we treat *one-sample's-V-modals-features* as *V-samples'-features*, and use PCA [61] to achieve the aligned features. The influence of the feature-alignment trick is shown in Table VII. Obviously, this trick is very helpful for our GCT, can improve its performance by about 1%-3%. (ii) Inspired by [62], we re-use the base data to fine-tune the frozen-pre-trained feature extractor with the smaller batchsize and learning rate to improve the model's performance. We fix the batchsize to 16 and the learning rate to 0.001. The influence of fine-tuning trick is also shown in Table VII. The results show that this trick can improve the performance of the original model by about 1%-2%.

TABLE IX

TIME CONSUMING. THE EXPERIMENTS ARE CONDUCTED IN OUR TESLA-V100 GPU. IN TESTING PHASE, THE REPORTED RESULTS ARE CONDUCTED IN THE INDUCTIVE SEMI-SUPERVISED SETTING WITH 15 UNLABELED SAMPLES. THE TESTING TIMES FOR DIFFERENT DATASETS ARE BASICALLY THE SAME BECAUSE WE SET THE EPISODES FOR EACH DATASET TO 600

Time (h)	mini-ImageNet	tiered-ImageNet	CIFAR-FS	FC100
Training	10.5	105.6	8.2	2.1
Fine-tuning	5.1	53.2	3.9	0.9
Testing (1-shot)	0.3	0.3	0.3	0.3
Testing (5-shot)	0.4	0.4	0.4	0.4



(a) 5-way 1-shot (b) 5-way 5-shot

Fig. 5. The effect of the number of unlabeled samples on IGL and GL. The experiments are conducted in mini-ImageNet with inductive semi-supervised setting without using the feature-alignment and fine-tuning tricks.

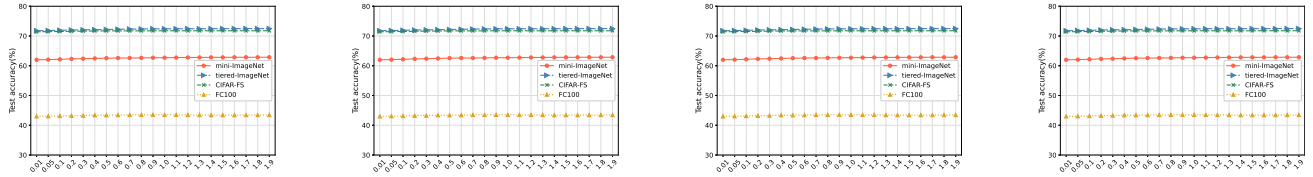
D. Comparison of GL and IGL

In this section, we will further analyze the GL and IGL in inductive semi-supervised setting. The experimental results are listed in Figure 5. We find that when there are fewer labeled samples, the performance of the GL method is slightly higher than the IGL method. But with the increase of unlabeled samples, we find that the performance of the GL method increases slower than our proposed IGL. The reason is that:

From the objective function of GL (i.e., Equation 18, we can directly get the solution as:

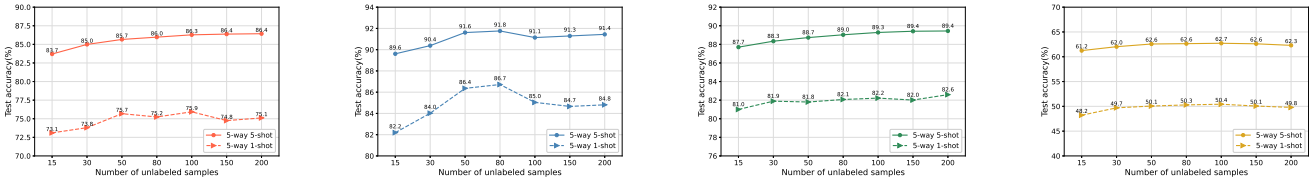
$$\mathbf{R} = (\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}})^{-1}\mathbf{Y} \tag{21}$$

where $\mathbf{R}$ is the to-be-calculated soft label matrix; $\mathbf{Y}$ indicates the initial label embedding matrix; $\Delta = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, $\mathbf{A}$ is the adjacency matrix, constructed by all the labeled and unlabeled samples.



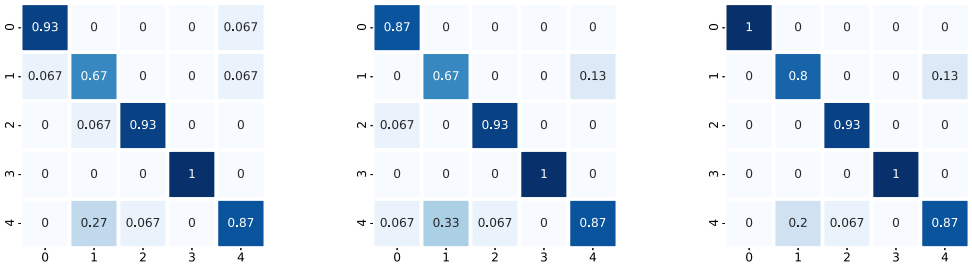
(a) λ (5-way 1-shot) (b) λ (5-way 5-shot) (c) μ (5-way 1-shot) (d) μ (5-way 5-shot)

Fig. 6. Influences of parameters. These experiments do not use the feature-alignment and fine-tuning tricks.



(a) mini-ImageNet (b) tiered-ImageNet (c) CIFAR-FS (d) FC100

Fig. 7. Comparison results of inductive semi-supervised few-shot classification with varied unlabeled samples. These experiments do not use the feature-alignment and fine-tuning tricks.



(a) SS-R-Mod (b) SS-M-Mod (c) Fusion-Mod

Fig. 8. Confusion matrices on mini-ImageNet.

While for the IGL, we predict the $\mathbf{x}_{ts}$'s category by Equation 10. Compared with Equation 21 and Equation 10, it's easy to find that: The sample feature has a far greater impact on Equation 10 than on Equation 21. That means, Equation 10 can better utilize the high-quality samples selected by the co-training strategy to get ideal results. To be more specific, in Equation 10, the selected samples will affect $\mathbf{X}$, $\mathbf{A}$ (when pseudo-labels are assigned to unlabeled samples through co-training, $\mathbf{A}$ will be reconstructed), and $\mathbf{Y}$. While in Equation 21, the selected samples will only affect $\mathbf{Y}$ (because all labeled samples and unlabeled samples are directly used when building $\mathbf{A}$ here, so $\mathbf{A}$ is fixed and will not be updated in the future). Therefore, in this case, the advantage of IGL over GL does not disappear. And on the contrary, the more unlabeled samples are used, the greater the improvement of IGL over GL.

E. Discussion About Self-Supervision Fusion

In this paper, we introduce two different self-supervised strategies (rotation-based and mirror-based) to strengthen the model and fuse them by our GCT in the meta-test phase. It is interesting to see the results directly fusing them into one branch in the pre-training stage. We list the results in Table VIII. Comparing ③ with ①, ②; ⑦ with ⑤, ⑥; we find that the results of fusing different self-supervision into one branch are better than the ones only using rotation-based or mirror-based strategy. But comparing ④ with ③; ⑧ with ⑦; we find that the experimental results obtained by GCT fusion in the meta-test stage are better than those obtained by directly fusing different self-supervision in the pre-training stage.

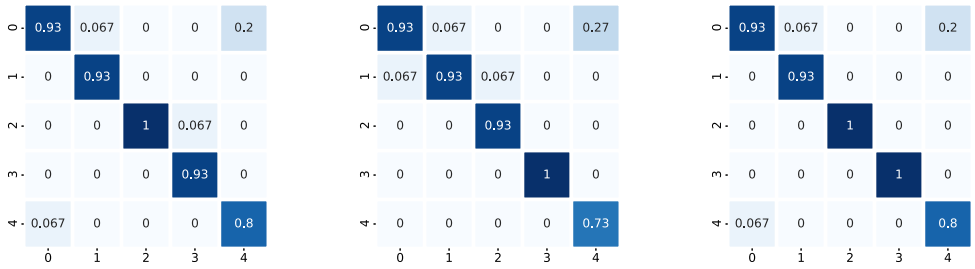
But this phenomenon doesn't mean that the co-training strategy is more effective than the self-supervised strategy. Their attention is different. The method based on self-supervision directly makes the pre-trained model have good transferability to extract more discriminative novel features, which is very helpful for cross-domain problems in few-shot classification tasks. However, the co-training method aims to strengthen the classifier to make it not be disturbed by noise features as much as possible. The two strategies are not a competitive relationship, but a cooperative relationship.

F. Time Consuming

In this section, we report the time-consuming of our method. It includes three parts, training time, fine-tuning time, and testing time. The results are listed in Table IX.

G. Multi-Modal Fusion

Besides SS-R-Mod and SS-M-Mod, we introduce Std-Mod and Meat-Mod (described in Section IV-E) to further evaluate the proposed method. We list the performances of more kinds of integrated ways on mini-ImageNet in Table X. All the results are on account of the inductive semi-supervised setting with 15 unlabeled samples. From the table, we can conclude that with the increase of fusion modalities, our method has the



(a) SS-R-Mod (b) SS-M-Mod (c) Fusion-Mod

Fig. 9. Confusion matrices on CIFAR-FS.

TABLE X

COMPARISON RESULTS OF FUSING MULTI-MODAL FEATURES IN 5-WAY 1-SHOT CASE. THE EXPERIMENTS ARE CONDUCTED ON THE INDUCTIVE SEMI-SUPERVISED SETTING WITH 15 UNLABELED SAMPLES. THESE EXPERIMENTS DO NOT USE THE FEATURE-ALIGNMENT FINE-TUNING TRICKS

Modal	1	2	3	4	5	6	7	8	9	10	11
Std-Mod	✓	✓	✓				✓	✓	✓		✓
Meta-Mod	✓			✓	✓		✓	✓		✓	✓
SS-R-Mod		✓		✓		✓	✓		✓	✓	✓
SS-M-Mod			✓		✓	✓		✓	✓	✓	✓
Accuracy	69.81	70.99	71.08	71.26	71.89	73.05	72.04	73.21	74.17	73.92	74.88

TABLE XI

COMPARISON IN CROSS-DOMAIN DATASET SCENARIO. OUR GCT IS ON INDUCTIVE SEMI-SUPERVISED SETTING WITH 15 UNLABELED SAMPLES. $(\cdot)^\flat$ AND $(\cdot)^\sharp$ REPRESENT THE REPORTED RESULTS COME FROM [63] AND [16], RESPECTIVELY. THESE EXPERIMENTS DO NOT USE THE FEATURE-ALIGNMENT AND FINE-TUNING TRICKS

Method	**mini-ImageNet ⟶ CUB**	
	1-shot	**5-shot**
Baseline$^\flat$ [47]	-	53.10
MatchingNet$^\flat$ [56]	-	53.10
MAML$^\flat$ [40]	-	51.30
ProtoNet$^\flat$ [39]	-	62.00
RelationNet$^\flat$ [41]	-	57.70
GNN$^\flat$ [64]	-	66.90
Neg-Cosine$^\flat$ [65]	-	67.00
LaplacianShot$^\flat$ [66]	-	66.30
TIM-GD$^\flat$ [63]	-	<u>71.00</u>
MetaOpt$^\sharp$ [33]	44.79	64.98
Manifold Mixup$^\sharp$ [67]	46.21	66.03
S2M2$^\sharp$ [16]	<u>48.24</u>	70.40
GCT	**59.55**	**76.07**

possibility of further improvement. And in this paper, we fix the fusion strategy to combine SS-R-Mod with SS-M-Mod for convenience.

H. Cross-Domain Few-Shot Learning

The GCT can be regarded as a highly robust approach in real-world situations, benefiting from the introduced multi-model information. That is, we estimate the proposed method with the transductive implements on a cross-domain dataset: *i.e.*, mini-ImageNet ⟶ CUB. The feature extractor was trained on the mini-ImageNet dataset in the pre-train process. And in the meta-test task, we classify the CUB dataset. All quantitative results are shown in Table XI. Compared with the SOTAs methods, our approaches have apparently improved, at least 11.3% in the 1-shot case and 5.1% in the 5-shot case. Based on the above experiment results, it is demonstrated that the GCT can solve the FEM problem better. The experimental results on the cross-domain few-shot learning task show that the proposed GCT would be effective in real situations.

VI. CONCLUSION

There is a fundamental problem in Few-shot learning based tasks, *i.e.*, Feature-Extractor-Maladaptive (FEM) problem. In this paper, we make two efforts to address this challenge. First, we propose a novel label prediction method, Isolated Graph Learning (IGL), to encode the feature embedding to graph representation and then propagate the label information through graph structure for prediction. Second, we extend IGL to the co-training framework to exploit multi-modal features in the semi-supervised setting, dubbed as Graph Co-Training (GCT). From the two perspectives, we have tackled this challenge to some extent. In our future work, we may study to improve the quality of the co-training strategy.

REFERENCES

[1] S. Shao, L. Xing, R. Xu, W. Liu, Y.-J. Wang, and B.-D. Liu, "MDFM: Multi-decision fusing model for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Dec. 16, 2021, doi: 10.1109/TCSVT.2021.3135023.

[2] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Dec. 2016.

[3] S. Shao, R. Xu, W. Liu, B.-D. Liu, and Y.-J. Wang, "Label embedded dictionary learning for image classification," *Neurocomputing*, vol. 385, pp. 122–131, Apr. 2020.

[4] L. Wang *et al.*, "Dense-scale feature learning in person re-identification," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 1–17.

[5] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2018.

[6] B. Fan *et al.*, "Contextual multi-scale feature learning for person re-identification," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 655–663.

[7] Y. Wang, Y. Zhao, S. Ying, S. Du, and Y. Gao, "Rotation-invariant point cloud representation for 3-D model recognition," *IEEE Trans. Cybern.*, early access, Mar. 22, 2022, doi: 10.1109/TCYB.2022.3157593.

[8] G. Qian, H. Hammoud, G. Li, A. Thabet, and B. Ghanem, "Assanet: An anisotropic separable set abstraction for efficient point cloud representation learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Dec. 2021, pp. 28119–28130.

[9] L. Zhao, K.-K. Ma, Z. Liu, Q. Yin, and J. Chen, "Real-time scene-aware LiDAR point cloud compression using semantic prior representation," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jan. 21, 2022, doi: 10.1109/TCSVT.2022.3145513.

[10] X. Li *et al.*, "Learning to self-train for semi-supervised few-shot classification," in *Proc. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 10276–10286.

[11] Y. Wang, C. Xu, C. Liu, L. Zhang, and Y. Fu, "Instance credibility inference for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12836–12845.

[12] K. Huang, J. Geng, W. Jiang, X. Deng, and Z. Xu, "Pseudo-loss confidence metric for semi-supervised few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8671–8680.

[13] M. Lazarou, T. Stathaki, and Y. Avrithis, "Iterative label cleaning for transductive and semi-supervised few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8751–8760.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Neural Inf. Process. Syst.*, 2002, pp. 585–591.

[15] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2003, pp. 321–328.

[16] P. Mangla, M. Singh, A. Sinha, N. Kumari, V. N. Balasubramanian, and B. Krishnamurthy, "Charting the right manifold: Manifold mixup for few-shot learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2218–2227.

[17] S. Shao *et al.*, "MHFC: Multi-head feature collaboration for few-shot learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 4193–4201.

[18] Y. Chen, Y. Ma, T. Ko, J. Wang, and Q. Li, "MetaMix: Improved meta-learning with interpolation-based consistency regularization," in *Proc. ICPR*, 2020, pp. 407–414.

[19] S. Huang, X. Zeng, S. Wu, Z. Yu, M. Azzam, and H.-S. Wong, "Behavior regularized prototypical networks for semi-supervised few-shot image classification," *Pattern Recognit.*, vol. 112, Apr. 2020, Art. no. 107765.

[20] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Neural Inf. Process. Syst.*, 2019, pp. 5049–5059.

[21] K. Sohn *et al.*, "FixMatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[22] Y. Lifchitz, Y. Avrithis, S. Picard, and A. Bursuc, "Dense classification and implanting for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9258–9267.

[23] N. Dvornik, C. Schmid, and J. Mairal, "Selecting relevant features from a multi-domain representation for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 769–786.

[24] L. Liu, W. Hamilton, G. Long, J. Jiang, and H. Larochelle, "A universal representation transformer layer for few-shot image classification," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–11.

[25] N. Dvornik, J. Mairal, and C. Schmid, "Diversity with cooperation: Ensemble methods for few-shot classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3723–3731.

[26] V. Garcia and J. Bruna, "Few-shot learning with graph neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

[27] J. Kim, T. Kim, S. Kim, and C. D. Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11–20.

[28] L. Yang, L. Li, Z. Zhang, X. Zhou, E. Zhou, and Y. Liu, "DPGN: Distribution propagation graph network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13390–13399.

[29] S. Tang, D. Chen, L. Bai, K. Liu, Y. Ge, and W. Ouyang, "Mutual CRF-GNN for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2329–2339.

[30] L. Yang, D. Chen, X. Zhan, R. Zhao, C. C. Loy, and D. Lin, "Learning to cluster faces via confidence and connectivity estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13369–13378.

[31] Z. Wang, L. Zheng, Y. Li, and S. Wang, "Linkage based face clustering via graph convolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1117–1125.

[32] Z. Zhang, H. Lin, X. Zhao, R. Ji, and Y. Gao, "Inductive multi-hypergraph learning and its application on view-based 3D object classification," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5957–5968, Dec. 2018.

[33] K. Lee, S. Maji, A. Ravichandran, and S. Soatto, "Meta-learning with differentiable convex optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10657–10665.

[34] M. Ren *et al.*, "Meta-learning for semi-supervised few-shot classification," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.

[35] Y. Liu *et al.*, "Learning to propagate labels: Transductive propagation network for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–14.

[36] P. Rodríguez, I. Laradji, A. Drouin, and A. Lacoste, "Embedding propagation: Smoother manifold for few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–138.

[37] Z. Yu, L. Chen, Z. Cheng, and J. Luo, "TransMatch: A transfer-learning scheme for semi-supervised few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12856–12864.

[38] H. Huang, J. Zhang, J. Zhang, Q. Wu, and C. Xu, "PTN: A Poisson transfer network for semi-supervised few-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 1602–1609.

[39] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4077–4087.

[40] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[41] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1199–1208.

[42] B. Oreshkin, P. R. López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 721–731.

[43] L. Qiao, Y. Shi, J. Li, Y. Tian, T. Huang, and Y. Wang, "Transductive episodic-wise adaptive metric for few-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3603–3612.

[44] J. Kim, H. Kim, and G. Kim, "Model-agnostic boundary-adversarial sampling for test-time generalization in few-shot learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 599–617.

[45] G. S. Dhillon, P. Chaudhari, A. Ravichandran, and S. Soatto, "A baseline for few-shot image classification," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–20.

[46] C. Simon, P. Koniusz, R. Nock, and M. Harandi, "Adaptive subspaces for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4136–4145.

[47] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and J.-B. Huang, "A closer look at few-shot classification," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.

[48] A. A. Rusu *et al.*, "Meta-learning with latent embedding optimization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–17.

[49] C. Xing, N. Rostamzadeh, B. Oreshkin, and P. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4847–4857.

[50] S. W. Yoon, J. Seo, and J. Moon, "TapNet: Neural network augmented with task-adaptive projection for few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7115–7123.

[51] H. Li, D. Eigen, S. Dodge, M. Zeiler, and X. Wang, "Finding task-relevant features for few-shot learning by category traversal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–10.

[52] C. Chen, K. Li, W. Wei, J. T. Zhou, and Z. Zeng, "Hierarchical graph neural networks for few-shot learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 240–252, Feb. 2021.

[53] S. Yang, L. Liu, and M. Xu, "Free lunch for few-shot learning: Distribution calibration," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–13.

[54] N. Fei, Z. Lu, T. Xiang, and S. Huang, "MELR: Meta-learning via modeling episode-level relationships for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–20.

[55] C. Xu *et al.*, "Learning dynamic alignment via meta-filter for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5182–5191.

[56] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 3630–3638.
[57] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200–2011 dataset," Univ. California, San Diego, CA, USA, Tech. Rep., 2011.
[58] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[59] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, USA, Tech. Rep., 2009.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[61] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 61, no. 3, pp. 611–622, 1999.
[62] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12664–12673.
[63] M. Boudiaf, Z. I. Masud, J. Rony, J. Dolz, P. Piantanida, and I. B. Ayed, "Transductive information maximization for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–13.
[64] H.-Y. Tseng, H.-Y. Lee, J.-B. Huang, and M.-H. Yang, "Cross-domain few-shot classification via learned feature-wise transformation," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–18.
[65] B. Liu *et al.*, "Negative margin matters: Understanding margin in few-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 438–455.
[66] I. Ziko, J. Dolz, E. Granger, and I. B. Ayed, "Laplacian regularized few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 11660–11670.
[67] V. Verma *et al.*, "Manifold mixup: Better representations by interpolating hidden states," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6438–6447.

Rui Xu received the B.S. and M.S. degrees from the College of Control Science and Engineering, China University of Petroleum, China, where she is currently pursuing the Ph.D. degree. Her main research interests include machine learning and computer vision.

Lei Xing received the B.S. degree from the College of Oceanography and Space Informatics, China University of Petroleum, China, where he is currently pursuing the M.S. degree. His main research interests include machine learning and computer vision.

Shuai Shao (Member, IEEE) received the M.S. degree from the College of Control Science and Engineering, China University of Petroleum, China, where he is currently pursuing the Ph.D. degree. He was a Visiting Student with Tsinghua University from 2019 to 2020. During his Ph.D. degree, he has published five articles as the first author in *ACM Multimedia* (ACMMM) and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT). His research interests include image processing, computer vision, and machine learning.

Lifei Zhao received the B.S. degree from the College of Control Science and Engineering, China University of Petroleum, China, where she is currently pursuing the M.S. degree.

Baodi Liu (Member, IEEE) received the B.S. degree in signal and information processing from the China University of Petroleum, China, in 2007, and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, in 2013. He was a Visiting Scholar with the University of California at Merced, Merced, from 2019 to 2020. He is currently an Associate Professor with the College of Control Science and Engineering, China University of Petroleum. His research interests include image processing, computer vision, and machine learning.

Weifeng Liu (Senior Member, IEEE) received the double B.S. degrees in automation and business administration and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2002 and 2007, respectively. He was a Visiting Scholar with the Centre for Quantum Computation and Intelligent Systems, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia, from 2011 to 2012. He is currently a Professor with the College of Control Science and Engineering, China University of Petroleum, China. His current research interests include pattern recognition and machine learning. He has authored or coauthored a dozen papers in top journals and prestigious conferences including ten ESI highly cited papers and three ESI hot papers. He serves as an Associate Editor for *Neural Processing Letters*; the Co-Chair for IEEE SMC Technical Committee on Cognitive Computing; and a Guest Editor for the special issues of *Signal Processing*, *IET Computer Vision*, *Neurocomputing*, and *Remote Sensing*. He also serves dozens of journals and conferences.

Yicong Zhou (Senior Member, IEEE) received the B.S. degree from Hunan University, Changsha, China, and the M.S. and Ph.D. degrees from Tufts University, MA, USA, all in electrical engineering.

He is currently a Professor and the Director of the Vision and Image Processing Laboratory, Department of Computer and Information Science, University of Macau. His research interests include image processing, computer vision, machine learning, and multimedia security. He is a fellow of the International Society for Optical Engineering (SPIE) and a Senior Member of the China Computer Federation (CCF). He was a recipient of the Third Price of Macao Natural Science Award in 2014 and 2020. He is the Co-Chair of Technical Committee on Cognitive Computing in the IEEE Systems, Man, and Cybernetics Society. He was listed as "World's Top 2% Scientists" on the Stanford University Releases List in 2020 and 2021 and the "Highly Cited Researcher" in the Web of Science in 2020 and 2021. He serves as an Associate Editor for IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON CYBERNETICS (TCYB), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), and four other journals.