DGIG-Net: Dynamic Graph-in-Graph Networks for Few-Shot Human–Object Interaction

Xiyao Liu, *Graduate Student Member, IEEE*, Zhong Ji, *Senior Member, IEEE*, Yanwei Pang, *Senior Member, IEEE*, Jungong Han, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Few-shot learning (FSL) for human–object interaction (HOI) aims at recognizing various relationships between human actions and surrounding objects only from a few samples. It is a challenging vision task, in which the diversity and interactivity of human actions result in great difficulty to learn an adaptive classifier to catch ambiguous interclass information. Therefore, traditional FSL methods usually perform unsatisfactorily in complex HOI scenes. To this end, we propose dynamic graph-in-graph networks (DGIG-Net), a novel graph prototypes framework to learn a dynamic metric space by embedding a visual subgraph to a task-oriented cross-modal graph for few-shot HOI. Specifically, we first build a knowledge reconstruction graph to learn latent representations for HOI categories by reconstructing the relationship among visual features, which generates visual representations under the category distribution of every task. Then, a dynamic relation graph integrates both reconstructible visual nodes and dynamic task-oriented semantic information to explore a graph metric space for HOI class prototypes, which applies the discriminative information from the similarities among actions or objects. We validate DGIG-Net on multiple benchmark datasets, on which it largely outperforms existing FSL approaches and achieves state-of-the-art results.**

***Index Terms*—Dynamic graph, few-shot learning (FSL), graph convolutional network (GCN), human–object interaction (HOI), metalearning.**

I. Introduction

UNDERSTANDING human actions and activities from vision information is a long-standing research for building an intelligent system [1]–[4]. One important direction is the human–object interaction (HOI), which aims at recognizing various relationships between human actions and surrounding objects. However, the development of the HOI study strikes a bottleneck, in which current techniques are difficult to address the imbalanced data distribution in HOI. Recently, few-shot learning (FSL) provides HOI a novel solution due to its potential to alleviate the low-data challenge.

FSL for HOI is proposed to recognize novel HOI categories effectively with a limited number of labeled examples [5]. It has the potential to address.

1) *Recognition of Tail Part in HOI Distribution:* HOI data are a natural long-tail distribution, where the instance imbalance among categories suffers from overfitting [6]. FSL methods learn a network that maps an unlabeled example (query sample) to its label from the small labeled support set [7], which imitates the capability of humans to identify objects with very little direct supervision.
2) *Combinatorial Explosion Problem in HOI:* Multiple labels in HOI cause the number of classes to increase exponentially, which results in difficulty to solve large-scale practical problems. FSL methods transfer the knowledge from existing HOI models to recognize novel visual concepts instead of training a new model from scratch. Although FSL methods provide the HOI scene with a promising direction, the HOI scene brings new challenges to the existing FSL methods.

The purpose of HOI emphasizes the relationships between objects and people, which are quite diverse and interactive. The same action with different objects is classified as different HOI categories, which results in different interclasses becoming ambiguous. For example, "Eat-Apple" and "Eat-Banana" are different classes in HOI, but they are similar in visual representations. This is a big challenge for current FSL methods, which are still in their infancy and only implemented in a simple and single scene, such as miniImageNet [7]. It is difficult to learn an adaptive learner for complex HOI scenes, which results in the unsatisfactory performance of FSL methods. An effective approach to improve the few-shot performance is to learn more representative and discriminative visual features, which provides sufficient evidence to perform classification with few samples.

For improving the representativeness and discriminability of instances, recent FSL studies develop two types of approaches. One is introducing auxiliary semantic modalities, such as label embeddings [5], [8]; attribute annotations [9]; and description text [10]. These approaches are inspired by language explanations that help infants recognize new visual objects,

Manuscript received June 29, 2020; revised October 9, 2020; accepted January 3, 2021. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771329 and Grant 61632018, and in part by the Central Funds Guiding the Local Science and Technology Development under Grant 206Z5001G. This article was recommended by Associate Editor S. Das. *(Corresponding author: Zhong Ji.)*

Xiyao Liu, Zhong Ji, and Yanwei Pang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China, and also with the Tianjin Key Laboratory of Brain-Inspired Intelligence Technology, Tianjin University, Tianjin 300072, China (e-mail: xiyaoliu@tju.edu.cn; jizhong@tju.edu.cn; pyw@tju.edu.cn).

Jungong Han is with the Computer Science Department, Aberystwyth University, Aberystwyth SY23 3FL, U.K. (e-mail: jungong.han@aber.ac.uk).

Xuelong Li is with School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China. (e-mail: xuelong_li@nwpu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3049537.

Digital Object Identifier 10.1109/TCYB.2021.3049537

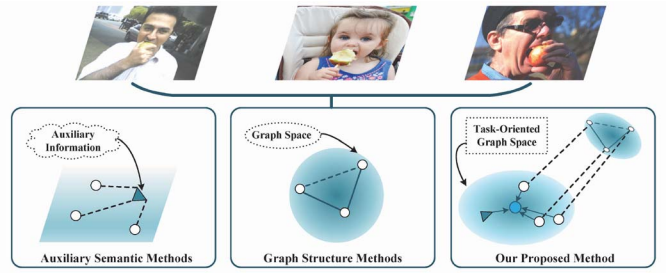


Fig. 1. Difference between our model and other methods. Auxiliary semantic methods introduce cross-modal information to help recognize new visual objects. Graph structure methods build a topological structure to improve the generalization of instance representations. Our proposed method designs a graph-in-graph structure to embed a visual subgraph to a dynamic graph-metric space guided by task-oriented semantic knowledge.

thus providing a strong information source for the data scarcity issue [11]–[13]. The other is designing a graph-based method due to its advantage on the effective representation of the graph-structured data. Graph structure models in the FSL have achieved promising results by dealing with complex relationships and interdependency among instances [14], [15]. Motivated by the above observations, we design a graph-in-graph structure to embed a visual subgraph to a dynamic graph-metric space guided by task-oriented semantic knowledge, as shown in Fig. 1.

To this end, we propose a dynamic graph-in-graph networks (DGIG-Net) for few-shot HOI by applying the dynamic and discriminative information from the similarities among actions or objects in this article, as shown in Fig. 2. It includes a knowledge reconstruction module (KR-Module) and a dynamic relation module (DR-Module), respectively. The KR-Module is designed to reconstruct the relationship among visual features to learn latent representations for HOI categories. Specifically, the encoder exploits both graph structure and node features with a graph convolutional network (GCN), and the decoder reconstructs the topological graph information and manipulates the latent graph representation. The DR-Module implements a graph metric space with dynamic task-oriented semantic information to obtain HOI class prototypes. It applies a cross-modal graph structure to encode two important types of knowledge: 1) the semantic guidance by action and object labels, dynamically defined by the label information from Word2Vector [16] and 2) the visual features obtained by the KR-Module.

It is worthwhile to highlight several aspects of the proposed approach here.

1) We implement a novel graph prototypes framework DGIG-Net by embedding a visual subgraph to a dynamic graph-metric space. In this way, it obtains HOI class prototypes instead of the linear prototypes method, which improves the representativeness and discriminability of the prototype features.
2) We design the KR-Module to encode both graph structure and node features with a GCN, and reconstruct the topological graph information, which manipulates the latent graph representation with a decoder. The KR-Module reconstructs the relationship among visual features to learn latent representations for HOI categories.
3) We develop a DR-Module that applies a cross-modal graph structure to encode dynamic semantic guidance by action and object labels and the visual features obtained by the KR-Module. The DR-Module implements a graph metric space with dynamic task-oriented semantic information to obtain HOI class prototypes.
4) Extensive experiments on two HOI benchmark datasets with two split strategies, that is, HICO-NN, TUHOI-NN, HICO-NF, and TUHOI-NF, demonstrate the effectiveness of our method. For example, our DGIG-Net improves accuracy by 3.7% in terms of 5-way 1-shot and 2.2% in terms of 5-way 5-shot on HICO-NF, and 5.4% and 2.8% on TUHOI-NF against the state-of-the-art methods, respectively. For the cross-domain few-shot HOI task, it also outperforms state-of-the-art methods.

The remaining sections of this article are organized as follows. Section II reviews the related work. Section III introduces our proposed DGIG-Net in detail. Section IV presents the experiments and analyses, followed by the conclusion in Section V.

II. Related Work

Our work is related to three active areas in machine learning: 1) FSL; 2) HOI recognition; and 3) GCN.

FSL: It is designed to train models for classification from only a handful of samples. There are three main types of methods to tackle the few-shot task: 1) metric-based; 2) optimization-based; and 3) generation-based approaches.

The methods in [7] and [17]–[21] aim at building metric-based networks by measuring the distance to realize FSL. For example, matching networks [7] apply a recurrent neural network (RNN) to accumulate task information in the embedding space of training samples to predict classes for testing samples. Remarkably, it defines the episode training strategy, which is widely applied by the following studies. Prototypical networks [17] learn a linear prototype space for classes and classify the query image into the nearest class prototypes. Relation networks [18] utilize neural networks to measure the possibility of two images belonging to the same class, which replaces the traditional artificial defining distance measurement method. DN4 [19] designs a local descriptor to learn the exchangeability of visual patterns across the images in the same class and complete image-to-class measurements. TPN [20] proposes learning a graph construction module to propagate labels from labeled instances to unlabeled test instances. Meanwhile, some studies explore more effective distance methods, such as [21]. It designs a two-stage few-shot approach to compute the Mahalanobis distances for class prototypes, which applies the features in low dimensionality to represent the relative relationship of samples.

Optimization-based FSL methods [22], [23] propose learning good initialization by adjusting the optimization algorithm and effectively obtain model parameters that can be learned with a few examples. For example, MAML [22] designs a model-agnostic method based on learning easily adaptable model parameters through gradient descent. Based on the idea, many methods extend this work, such as Reptile [23]

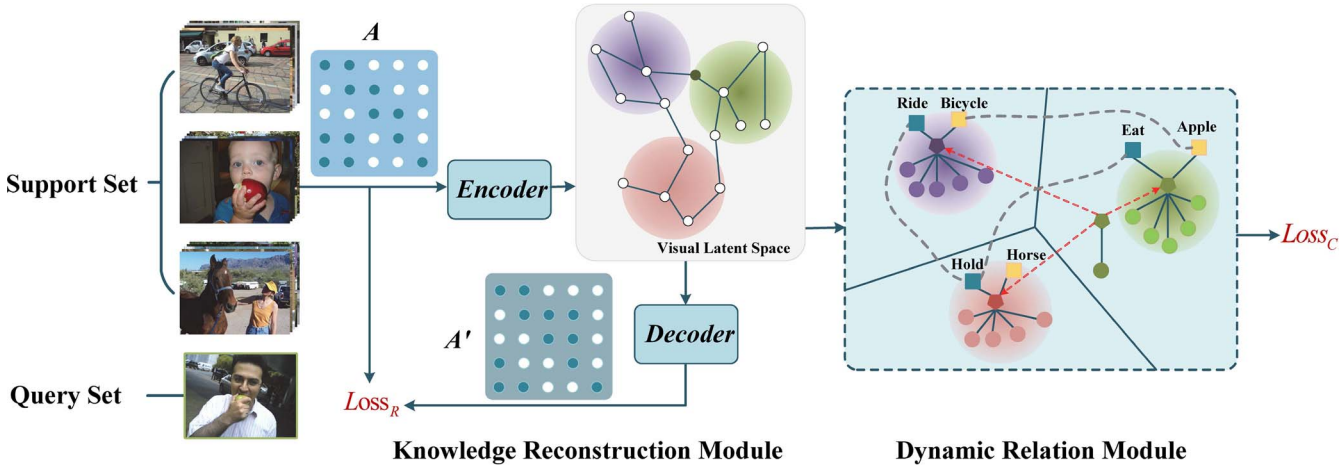


Fig. 2. Proposed DGIG-Net architecture, taking the 3-way 5-shot task for example. Given an adjacency matrix A, the KR-Module learns a latent representation through a graph encoder and then manipulates the input of the DR-Module, which applies a cross-modal graph structure to obtain HOI class prototypes with dynamic task-oriented semantic information.

and LEO [24]. For example, Reptile [23] directly implements stochastic gradient descent (SGD) in training instead of computing twice gradients, which requires less computation than MAML.

Moreover, some work addresses the few-shot problem with a data-driven solution, called generation-based methods. These methods create novel samples to augment the training set and improve the performance of current few-shot algorithms. Wang *et al.* [25] trained a generative adversarial network (GAN) to generate new instances, which achieves up to a 6% boost in classification accuracy when only given a single training example. Zhang *et al.* [26] replaced backgrounds and foregrounds guided by saliency maps to generate new support samples, which is more effective and less costly than GAN.

Few-shot HOI is a challenging vision task due to the fact that it is difficult to learn an adaptive classifier for complex HOI scenes. For improving the representativeness and discriminability of instances in HOI, we propose a DGIG-Net, a novel graph prototypes framework to learn a dynamic graph-metric space guided by a task-oriented semantic for few-shot HOI.

HOI Recognition: As a subtask of human action recognition, HOI recognition is introduced for alleviating such ambiguities caused by no motion cue in a still image. Since actions involving objects provide a context in spatial and functional relations, human behaviors are recognized effectively [27]. Early research relied on shape features and movement analysis to recognize HOI [28].

Recently, deep learning technologies have brought promising results on the HOI recognition task. Successively, large-scale image datasets [6], [29] were also released. Some work found that humans interact with an object by contacting some parts of the body instead of the entire body. Gkioxari *et al.* [30] developed a part-based model to make fine-grained action recognition based on the input of both whole person and part bounding boxes. Fang *et al.* [31] proposed a new pairwise body-part attention model that can learn to focus on crucial parts and their correlations for HOI recognition. Moreover, Mallya and Lazebnik [32] attached importance to HOI in visual question answering (VQA), and proposed a deep convolutional network model that fuses features from local and global context to recognize HOI.

Some studies focus on instance-based HOI detection tasks [33], [34], which utilize holistic human poses and global context under the help of popular detectors jointly to infer the locations and categories of HOI. In addition, compositional learning [35] employs an external knowledge graph to recognize unseen interactions, which applies zero-shot learning to address the data scarcity problem in HOI. To further alleviate the instance imbalance and combinatorial explosion challenges in HOI recognition, SGAP-Net [5] formulates HOI as a few-shot HOI task and learns a semantic-guided metric space to obtain attentive class prototypes for few-shot HOI.

GCN: Traditional neural networks have made great progress in Euclidean data but still perform unsatisfactorily on non-Euclidean domains. Graph neural networks raise attention by dealing with complex relationships and interdependency among instances [36]. Several studies have applied different types of graph neural networks in node classification [35], link prediction [37], and graph classification [38]. For example, Kato *et al.* [35] proposed a zero-shot HOI method, which constructs an external knowledge graph and GCN to recognize novel action–object compositions in HOI. Qi *et al.* [37] inferred a parse graph neural network that includes the HOI graph structure represented by an adjacency matrix, and the node labels for HOI detection. Mallea *et al.* [38] proposed a model for graph classification by extracting fixed size tensorial information from each graph in a given set, and employing a capsule network to perform classification.

GCN is widely applied in node regression and node classification tasks. Kipf and Welling [39] proposed GCN to formulate semisupervised classification as graph node classification. Incremental improvements [40]–[42] have been made over GCN. For example, adaptive GCN (AGCN) [40] is not restricted on graph degree by employing unspecified generalized and flexible structural relations. Dual GCN (DGCN) [41] proposes a DGCN architecture embedding semantic information with two graph convolutional layers in parallel.

Recently, graph-based methods have been employed in FZL [43]–[46]. For example, Gidaris and Komodakis [44] designed a GCN-based denoising autoencoder network by taking as input a set of classification weights corrupted with Gaussian noise to reconstruct the target-discriminative classification weights, which regularizes the weight generating metamodel. Iscen *et al.* [45] focused on modeling clean and noisy data by a graph per class and predicting class relevance of noisy examples. Although GCN has been directly applied in a number of recent FSL methods, the difference of our DGIG-Net lies in that we apply GCN to explore the cross-modal relationship among semantic information, class prototypes, and visual instances, which learns a graph prototypes metric space to obtain HOI prototypes.

III. DGIG-Net for Few-Shot HOI Recognition

In this section, we present our proposed DGIG-Net. The architecture of DGIG-Net consists of a KR-Module and a DR-Module, as shown in Fig. 2. We first develop a graph autoencoder KR-Module, which effectively applies both interclass and intraclass information to learn a latent representation. Based on the representation, a DR-Module is then proposed to integrate category semantic information and visual information toward cross-modal dynamic prototypes. We first introduce the problem formulation and then report our approach in detail.

A. Problem Definition

In few-shot classification, there is a metatrain set $\mathcal{S} = \{x_i, l_i, y_i\}_{i=1}^{N}$ that consists of N samples from M categories, where each $x_i \in \mathbb{R}^D$ is a D-dimensional visual feature vector of the ith image, l_i is its label semantic embeddings, and y_i is the one-hot class label. According to the datasets of HOI, the label for an image combines a pair of the action and the object. Every sample in the metatrain set is randomly divided into the support set or the query set. When training in the support set, the semantic vectors of labels are given as $l_i = \{(n_i, v_i)\}$, where $n_i \in \mathbb{R}^V$ is the V-dimensional text semantic embedding of the noun label and $v_i \in \mathbb{R}^V$ is the V-dimensional text semantic embedding of the verb label. There are no semantic labels in the query set.

We follow a C-way K-shot episode-based training strategy defined by matching networks [7]. Each episode is formed by sampling C classes and K labeled samples of each class from $\mathcal{S}$ to construct a few-shot task, which contains a support set and a query set to simulate the training and testing process. Specifically, in the wth episode, the support set can be denoted as $\mathcal{S}^w_{\text{support}} = \{x_i, l_i, y_i\}_{i=1}^{N_s} (N_s = C \times K)$, and the query set $\mathcal{S}^w_{\text{query}} = \{x_i, y_i\}_{i=1}^{N_q}$.

B. Knowledge Reconstruction Module

To represent the relationship of visual space and graph structure in a unified framework, we develop a new graph autoencoder network as a graph encoder. The idea is to learn the hidden representations of each node by combining support samples of the same categories, and to integrate interclass visual features with the graph structure in the latent representation. The most straightforward strategy to attend the neighbors of a node is to embed its representation from all its neighbors.

Formally, we define our knowledge reconstruction graph as $\mathcal{G} = (\mathcal{V}, \mathcal{Z})$. $\mathcal{G}$ is an undirected graph with $\mathcal{V}$ as its nodes. $\mathcal{Z}$ are the feature vectors for nodes $\mathcal{V}$. Specifically, we deploy an adjacency matrix to represent the visual relationship between support and query samples

$$A_{KR} = \begin{bmatrix} A_{ss} & 0 \\ 0 & A_{qq} \end{bmatrix}, \quad Z_{KR} = \begin{bmatrix} Z_s, & Z_q \end{bmatrix} \tag{1}$$

where A_{KR} and Z_{KR} are the adjacency matrix and node features of the knowledge reconstruction graph, A_{ss} and A_{qq} are adjacency matrices for support–support nodes and query–query nodes, and Z_s and Z_q are visual features of the support set and the query set, respectively. This graph includes 2 types of nodes: 1) support nodes and 2) query nodes. The adjacency matrix A_{ss} is defined as

$$A_{ss}(i,j) = \begin{cases} 1, & \text{if } y_i = y_j \text{ and } i! = j \\ 0, & \text{else} \end{cases} \tag{2}$$

where y_i is the label of the ith support sample. $A_{qq} = 0$, since query nodes only have self-connection to update its node features, which will be added after normalization. To better capture the graph structure, the adjacency matrix is normalized as being real symmetric positive semidefinite [39]. The adjacency A_{KR} is normalized as

$$\widehat{A}_{KR} = D^{-\frac{1}{2}}(A_{KR} + I)D^{-\frac{1}{2}} \tag{3}$$

where D is the diagonal node degree matrix, and I is an identity matrix to add self-connection to each node. The structure in our adjacency matrix could be denoted as

$$\widehat{A}_{KR} = \begin{bmatrix} \widehat{A}_{ss} & 0 \\ 0 & \widehat{A}_{qq} \end{bmatrix} \tag{4}$$

where $\widehat{A}_{ss}$ and $\widehat{A}_{qq}$ are adjacency matrices for support–support nodes and query–query nodes after graph normalization.

All visual node features Z are obtained by transforming the node features that they link on the graph in GCN. Formally, a single-layer GCN is calculated as

$$\widetilde{Z} = \text{GCN}(Z, A) = \widehat{A}Z^T W \tag{5}$$

where $\widehat{A}$ is the normalized graph adjacency matrix, Z is the node features, and W is the training weight parameter matrix. GCN transforms features on each node guided by the adjacency matrix. This operation is usually stacked with multilayer, where nonlinear activation functions (i.e., ReLU) are applied, as shown in Fig. 3.

Since $\widehat{A}$ is a block matrix, it can further decompose each GCN layer to each block. This decomposition provides better insights for our model. Specifically, we have

$$\begin{aligned} \widetilde{Z}_s &= \widehat{A}_{ss} Z_s W \\ \widetilde{Z}_q &= \widehat{A}_{qq} Z_q W \end{aligned} \tag{6}$$

where $\widetilde{Z}_s$ and $\widetilde{Z}_q$ are the outputs of our knowledge reconstruction graph that can also be noted by $Z_v = [\widetilde{Z}_s, \widetilde{Z}_q]$. There are various types of decoders, which reconstruct either the

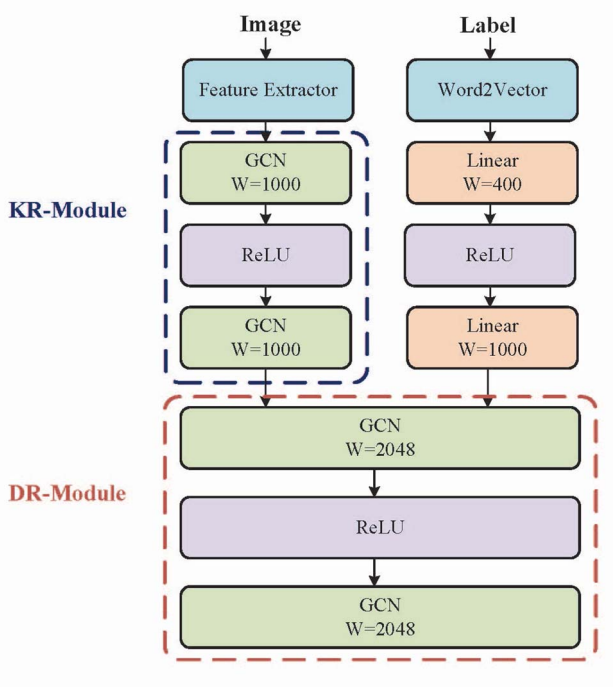


Fig. 3. Detailed network architectures applied in DGIG-Net.

graph structure or the attribute value [47]. As our latent embedding already contains both content and structure information, we apply a simple inner product decoder to predict the links between nodes, which would be efficient and flexible

$$A' = \text{Sigmoid}\left(Z_v^T Z_v\right) \tag{7}$$

where A' is the reconstructed structure matrix in the graph. We minimize the reconstruction error by measuring the difference between A and A'

$$\text{Loss}_R = KL\left(A'||A\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} A'_{ij} \log \frac{A'_{ij}}{A_{ij}}. \tag{8}$$

C. Dynamic Relation Module

One of the main challenges for few-shot methods is the dynamic adaption for different tasks. To confront this challenge, we develop a dynamic relation graph algorithm as the solution, which generates a task-oriented graph prototypes metric space.

We design a cross-modal graph structure to encode two important types of knowledge: 1) the semantic guidance by the verb and noun labels, defined by the label information from Word2Vector [16] and 2) the visual features obtained by the KR-Module. Especially, a dynamic graph adjacency matrix is applied in our graph structure, which is guided by task-oriented semantic information. It generates different adjacency matrices by action similarity and object similarity for different tasks. The dynamic adjacency matrices could transform node features by borrowing information from similar actions and objects. For example, when "Eat-Apple" and "Ride-Bicycle" appear in a task, our adjacency matrix will calculate the similarity of "Eat" and "Ride," and "Apple" and "Bicycle," as shown in Fig. 4. Then, the adjacency matrix links the two semantic nodes by their word similarity. Thus, the entire graph is dynamic for different tasks.

Graph Construction: Specifically, we construct the graph as follows.

1) There are three types of nodes in our graph. These nodes are denoted as: 1) label semantic nodes; 2) class prototype nodes; 3) and visual feature nodes, where their node features are denoted as Z_l, Z_p, and Z_v, respectively.
2) Each class prototype node defines an HOI class prototype. These class prototypes are modeled by a separate set of label semantic nodes Z_l and visual nodes Z_v in the graph. These nodes are initialized with all zero feature vectors and will obtain their representations Z_p via integrating category semantic information and visual information.
3) A label semantic (verb or noun) node only connects to a class prototype node. Similarly, each visual (support or query) node only links to a class prototype node.
4) WordNet [48] is applied to create noun–noun and verb–verb links, which generates dynamic relationships for the entire graph by the multilayer GCN encoder.

The graph construction of the DR-Module is shown in Fig. 4. This graph is thus captured by its adjacency matrix $A \in \mathbb{R}^{|V|\times|V|}$ and a feature matrix $Z \in \mathbb{R}^{d\times|V|}$. Based on the construction, our graph structure can be naturally decomposed into blocks, given by

$$A_{DR} = \begin{bmatrix} A_{ll} & A_{lp} & 0 \\ A_{lp}^T & A_{pp} & A_{pv} \\ 0 & A_{pv}^T & A_{vv} \end{bmatrix}, \quad Z_{DR} = \begin{bmatrix} Z_l, & Z_p, & Z_v \end{bmatrix} \tag{9}$$

where A_{ll}, A_{lp}, A_{pp}, A_{pv}, and A_{vv} are adjacency matrices for label–label pairs, label–prototype pairs, prototype–prototype pairs, prototype–visual pairs, and visual–visual pairs, respectively. Z_l, Z_p, and Z_v are node features. Especially, a generation network is applied to embed semantic vectors to visual features space to obtain Z_l, which is a 2-layer neural network with ReLU function. Moreover, we have $Z_p = 0$ and, thus, the prototypes nodes need to learn new representations for recognition. A_{ll} is the adjacency matrix containing the relationship between action label nodes and object label nodes, which is demoted as

$$A_{ll} = \begin{bmatrix} A_{aa} & A_{ao} \\ A_{ao}^T & A_{oo} \end{bmatrix} \tag{10}$$

where A_{aa}, A_{ao}, and A_{oo} are adjacency matrices for action–action label pairs, action–object label pairs, and object–object label pairs, respectively. A_{ll} is defined by the similarity of words calculated by WordNet [48], which is denoted as

$$A_{ll}(i,j) = \begin{cases} \text{Object Similarity}, & \text{if } y_i, y_j \in \text{ Object} \\ \text{Action Similarity}, & \text{if } y_i, y_j \in \text{ Action} \\ 0, & \text{else} \end{cases} \tag{11}$$

where Object Similarity and Action Similarity are obtained by WordNet [48] for each task. Thus, A_{ll} is dynamically decided by different tasks, which makes the DR-Module dynamic. A_{lp}, A_{pp}, A_{pv}, and A_{vv} are defined similarly with (2).

Similarly, the adjacency matrix A_{DR} needs to be normalized by (3). The structure in our adjacency matrix could be

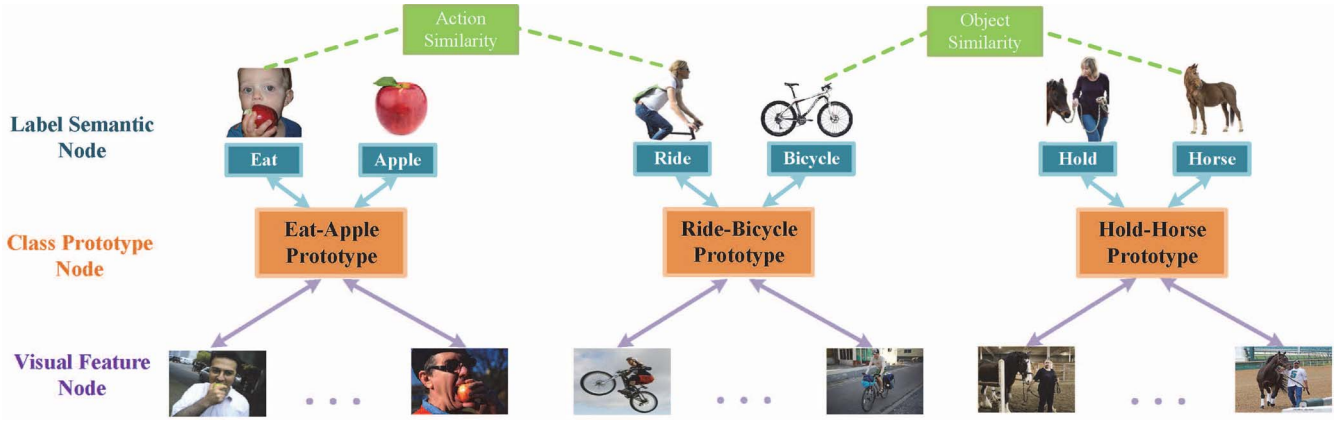


Fig. 4. Illustration of the DR-Module. There are three types of nodes in this graph, which are denoted as label semantic nodes, class prototype nodes, and visual feature nodes, respectively. The label semantic nodes provide dynamic relationships from their corresponding task, which are integrated into class prototype nodes by GCN encoder.

transformed as

$$\widehat{A}_{DR} = \begin{bmatrix} \widehat{A}_{ll} & \widehat{A}_{lp} & 0 \\ \widehat{A}_{lp}^T & \widehat{A}_{pp} & \widehat{A}_{pv} \\ 0 & \widehat{A}_{pv}^T & \widehat{A}_{vv} \end{bmatrix} \tag{12}$$

where $\widehat{A}_{pp}$, $\widehat{A}_{pv}$, and $\widehat{A}_{vv}$ are adjacency matrices for prototype–prototype nodes, prototype–visual nodes, and visual–visual nodes after graph normalization.

The prototype node features Z_p are obtained by transforming the node features that they link on the graph in GCN. Formally, a single-layer GCN is calculated by (5).

After the first layer GCN encoder, we have

$$\begin{aligned} \widetilde{Z}_l &= \left(\widehat{A}_{ll} Z_l + \widehat{A}_{lp} Z_p\right) W \\ \widetilde{Z}_p &= \left(\widehat{A}_{lp}^T Z_l + \widehat{A}_{pp} Z_p + \widehat{A}_{pv} Z_v\right) W \\ \widetilde{Z}_v &= \left(\widehat{A}_{vv} Z_v + \widehat{A}_{pv} Z_p\right) W \end{aligned} \tag{13}$$

where $\widetilde{Z}_l$, $\widetilde{Z}_p$, and $\widetilde{Z}_v$ are the outputs of our HOI graph that can also be noted by $\widetilde{Z} = [\widetilde{Z}_l, \widetilde{Z}_p, \widetilde{Z}_v]$. With nonlinear activations and multilayer GCN, the model will construct a nonlinear transform that considers more nodes for building the HOI class prototypes. We implement the output HOI prototype representations $\widetilde{Z}_p$ for the HOI class prototypes in few-shot HOI.

We calculate a probability distribution by the distance between a query sample and the class prototypes of support set to accomplish the recognition task

$$p_\phi\left(y = c | q \in \mathcal{S}^w_{\text{query}}\right) = \frac{\exp\left(-d\left(z_p^q, z_p^c\right)\right)}{\sum_k \exp\left(d\left(z_p^q, z_p^k\right)\right)} \tag{14}$$

where $d(\cdot)$ is the Euclidean distance, z_p^c is the prototype features of class c in $\widetilde{Z}_p$, and z_p^q is the query q representation after GCN in $\widetilde{Z}_p$.

Besides, in the training process, we apply a cross-entropy loss to measure the classification error

$$\text{Loss}_C = d\left(z_p^q, z_p^c\right) + \log \sum_{N_C} \exp\left(-d\left(z_p^q, z_p^c\right)\right). \tag{15}$$

Thus, the final loss of the entire DGIG-Net consists of two parts: 1) reconstruction loss and 2) classification loss, which is denoted as

$$\text{Loss} = \text{Loss}_C + \lambda \text{Loss}_R \tag{16}$$

where λ is a hyperparameter adjusting the weight of the two modules.

IV. Experiments

A. Experiment Setup

The CNN structure of our model is a pretrained ResNet-18 [49]. Thus, each input image is represented as a 1000-D vector. The Adam optimizer is utilized with the initial learning rate of 0.000001. The hyperparameter λ is set to be 0.1. In terms of the regularizer, we set 0.01 for all datasets. Besides, we apply Word2vector [16] to extract the semantic embeddings for the category labels, which are represented as 400-D vectors.

B. Datasets

Among the widely available datasets for HOI, we select two popular datasets, namely: 1) humans interacting with common objects (HICOs) [6] and 2) Trento Universal HOI (TUHOI) [29]. The HICO dataset consists of 42 109 images with 80 objects and 92 actions, which covers almost human daily activities with 377 interactions. For establishing a more natural and realistic dataset, TUHOI collects images first and then defines actions from images instead of some predefining human actions. Thus, TUHOI is a small scale but rich dataset, which contains 9802 images with 95 objects, 66 actions, and 194 interactions.

To satisfy the need for our experiments, original datasets should be divided into novel compositions. Following the popular setting of FSL [7], [17], we apply 60/20/20 training/validation/testing repartitions for reorganizing the datasets. We present two split strategies: 1) novel noun (NN) and 2) novel few instances (NF). Details of both strategies are described as follows.

1) Novel Noun: This is the first split strategy. We follow that ubiquitous similarity exists in the same action interacting

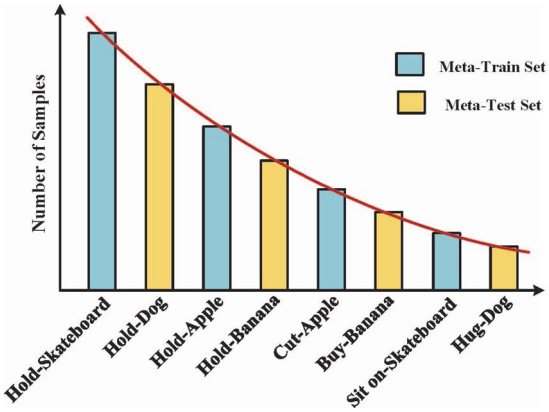


Fig. 5. NN-datasets divide objects as different tasks. The long-tail distribution with disjoint nouns is preserved in each set (taking HICO-NN as an example).

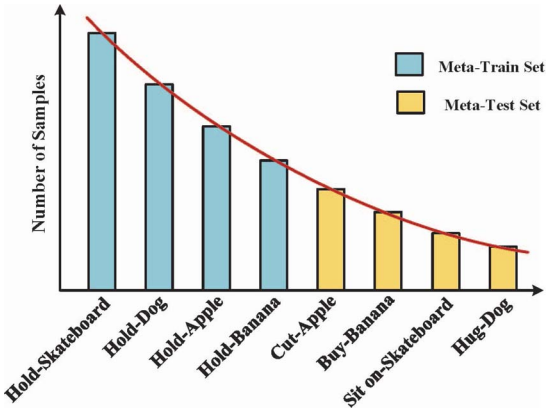


Fig. 6. NF-datasets divide categories with more samples as the metatrain and others are employed in the metatest set and the metavalidation set. The head part of the long-tail distribution appears in the metatrain and the metatest set presents the tail part (taking HICO-NF as an example).

TABLE I
SETTINGS OF NN-DATASETS

Dataset	Item	Meta-train Set	Meta-Val Set	Meta-Test Set	Total
HICO-NN	Action	69	42	47	-
	Object	45	15	20	80
	Interaction	212	73	92	377
	Image	24,067	8,896	9,146	42,109
TUHOI-NN	Action	46	24	24	-
	Object	50	20	25	95
	Interaction	98	44	52	194
	Image	4,871	2,361	2,570	9,802

TABLE II
SETTINGS OF NF-DATASETS

Dataset	Item	Meta-train Set	Meta-Val Set	Meta-Test Set	Total
HICO-NF	Action	50	54	49	-
	Object	65	61	59	-
	Interaction	173	102	102	377
	Image	38,147	1,987	1,975	42,109
TUHOI-NF	Action	30	32	28	-
	Object	58	53	53	-
	Interaction	69	62	63	194
	Image	7,294	1,277	1,231	9,802

with different objects. Moreover, objects could guide a set of behaviors, that is, apple can be eaten, held, etc. Similar actions with different objects can be transferable knowledge. Thus, we divide objects as different tasks in our work. Specifically, we divide all noun labels into the metatrain set, the metaval set, and the metatest set that are disjoint in nouns. For example, the object "apple" only appears in the metatrain set and corresponding similar object "banana" is divided into the metatest set, as shown in Fig. 5.

HICO-NN: The modified dataset HICO divided with the NN strategy is called HICO-NN. We divide HICO-NN into a metatrain set with 45 nouns and 24 067 images, a metatest set with 20 nouns and 9146 images, and a metavalidation set with 15 nouns and 8896 images, which are disjoint in noun labels.

TUHOI-NN: TUHOI-NN is reorganized similarly to that of the HICO-NN dataset. There are 50 nouns and 4871 images in the metatrain set, 20 nouns and 2361 images in the metaval set, and 25 nouns and 2570 images in the metatest set of TUHOI-NN. More details are listed in Table I.

2) Novel Few Instances: This is the second split strategy. Since the data show a long tail distribution and our purpose is to recognize the unusual categories, we select the categories that have over 50 samples as the metatrain, and others are employed in the metatest set and metavalidation set, as shown in Fig. 6.

HICO-NF: We call the modified dataset HICO for NF instances as HICO-NF. There are 173 interactions and 38 147 images in the metatrain set, 102 interactions and 1987 images in the metavalidation set, and 102 interactions and 1975 images in the metatest set. The categories in the metavalidation and metatest set contain 49 samples at most and 6 samples at least.

TUHOI-NF: Similar to that of HICO-NF dataset, we reorganize TUHOI to be TUHOI-NF. We divide it into a metatrain set with 69 interactions and 7294 images, a metavalidation set with 62 interactions and 1277 images, and a metatest set with 63 interactions and 1231 images. More details are listed in Table II.

C. Comparison With State-of-the-Art Methods

We compared a total of eight few-shot approaches with our model in our experiments. These few-shot algorithms include metric-based and optimization-based methods as follows.

Metric-Based Methods: Matching networks [7] apply an RNN to accumulate task information in the embedding space.

Prototypical networks [17] train a CNN to embed task examples to a metric space and perform nearest neighbor classification with the class prototypes.

Relation networks [18] introduce a learnable metric network to compare the similarity of different samples.

DN4 [19] designs a local descriptor to learn the exchangeability of visual patterns across the images in the same class, and complete-based image-to-class measures.

TPN [20] proposes learning a graph construction module to propagate labels from labeled instances to unlabeled test instances.

TABLE III
FEW-SHOT CLASSIFICATION ACCURACY OF DGIG-NET ON HICO-NN AND TUHOI-NN WITH ± 95% CONFIDENCE INTERVALS

Method	Type	HICO-NN		TUHOI-NN	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Matching Networks [7]	Metric	32.14 ± 1.62%	44.87 ± 1.74%	32.48 ± 1.58%	40.04 ± 1.70%
Prototypical Networks [17]	Metric	32.56 ± 1.59%	42.49 ± 1.75%	31.12 ± 1.55%	39.26 ± 1.71%
Relation Networks [18]	Metric	33.20 ± 1.68%	46.15 ± 1.81%	33.50 ± 1.68%	41.15 ± 1.75%
DN4 [19]	Metric	33.07 ± 1.43%	46.19 ± 1.74%	32.49 ± 1.43%	41.75 ± 1.77%
TPN [20]	Metric	33.40 ± 1.55%	46.33 ± 1.86%	32.95 ± 1.59%	41.73 ± 1.79%
EGNN [46]	Metric	34.25 ± 1.58%	48.12 ± 1.76%	33.76 ± 1.64%	43.48 ± 1.77%
SGAP-Net [5]	Metric	38.16 ± 1.65%	58.39 ± 1.82%	37.27 ± 1.61%	57.05 ± 1.73%
MAML [22]	Optimization	33.87 ± 1.74%	47.25 ± 1.84%	33.78 ± 1.64%	43.67 ± 1.79%
Reptile [23]	Optimization	33.26 ± 1.77%	46.56 ± 1.85%	32.39 ± 1.81%	41.65 ± 1.93%
DGIG-Net (Ours)	Metric	**39.13 ± 1.68%**	**59.06 ± 1.89%**	**38.77 ± 1.49%**	**58.07 ± 1.89%**

TABLE IV
FEW-SHOT CLASSIFICATION ACCURACY OF DGIG-NET ON HICO-NF AND TUHOI-NF WITH ± 95% CONFIDENCE INTERVALS

Method	Type	HICO-NF		TUHOI-NF	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Matching Networks [7]	Metric	34.95 ± 1.64%	46.01 ± 1.70%	33.00 ± 1.63%	41.59 ± 1.72%
Prototypical Networks [17]	Metric	34.88 ± 1.60%	45.47 ± 1.76%	32.04 ± 1.57%	41.27 ± 1.75%
Relation Networks [18]	Metric	36.62 ± 1.68%	48.01 ± 1.89%	36.05 ± 1.67%	42.35 ± 1.73%
DN4 [19]	Metric	35.21 ± 1.43%	47.36 ± 1.74%	35.47 ± 1.43%	44.72 ± 1.77%
TPN [20]	Metric	36.24 ± 1.37%	49.35 ± 1.66%	37.67 ± 1.56%	45.32 ± 1.67%
EGNN [46]	Metric	39.22 ± 1.59%	54.13 ± 1.79%	38.55 ± 1.58%	52.93 ± 1.75%
SGAP-Net [5]	Metric	43.37 ± 1.66%	70.78 ± 1.81%	41.12 ± 1.64%	69.47 ± 1.74%
MAML [22]	Optimization	38.86 ± 1.69%	53.32 ± 1.90%	36.45 ± 1.84%	48.48 ± 1.90%
Reptile [23]	Optimization	38.49 ± 1.69%	52.98 ± 1.78%	37.26 ± 1.83%	49.29 ± 1.88%
DGIG-Net (Ours)	Metric	**47.08 ± 1.44%**	**73.06 ± 1.62%**	**46.54 ± 1.49%**	**72.36 ± 1.62%**

EGNN [46] employs a deep neural network on the updates of both node features and edge features for FSL.

SGAP-Net [5] is the first approach designed for few-shot HOI, which learns a semantic-guided metric space to obtain attentive class prototypes.

Optimization-Based Methods: MAML [22] provides a parameter-optimization method for an arbitrary learner model that can be quickly adapted to a particular task.

Reptile [23] generalizes first-order MAML and ignores second-order derivatives, which requires less computation and memory than MAML.

These approaches all utilize ResNet-18 [49] as the embedding network. It is computed by averaging ten times over 600 randomly generated episodes as few-shot HOI recognition accuracy.

1) Comparison on NN Split Strategy: Table III describes the classification performance of DGIG-Net and eight competitors on HICO-NN and TUHOI-NN. We observe that our approach beats the state of the art in terms of both 5-way 5-shot and 5-way 1-shot tasks on both datasets. Specifically, compared with the second-best method SGAP-Net, the accuracy improvement on HICO-NN in terms of 5-way 1-shot increases from 38.16% to 39.13%, and in terms of 5-way 5-shot from 58.39% to 59.06%. On TUHOI-NN datasets, the proposed DGIG-Net also gains improvements from 37.27% to 38.77% in terms of 5-way 1-shot and from 57.05% to 58.07% in terms of 5-way 5-shot, which outperform the state-of-the-art approaches at least in 1.5% and 1.0%. Moreover, compared with the metric-based approaches, our DGIG-Net achieves obvious improvements on both datasets, which demonstrates that graph prototypes capture a more discriminative metric space than the others. Compared with the optimization-based methods, our DGIG-Net has more significant improvements, which indicates that our model with task-oriented dynamic relation is more transferable than learning the optimization strategy.

We also observe that the results of all methods on TUHOI-NN are lower than those on HICO-NN. We suppose the reason lies in the original distribution of data: the average samples of TUHOI-NN are much less than those of HICO-NN. Therefore, the few-shot HOI task is more difficult on TUHOI than that on HICO. Remarkably, DGIG-Net achieves 38.77% in terms of 5-way 1-shot on TUHOI-NN datasets, which significantly outperforms the second-best performance by 1.5%. It demonstrates that the dynamic relation graph structure of DGIG-Net has the superior ability on the difficult dataset with fewer samples. Moreover, our work applies task-oriented semantic guidance to capture class discriminative information, which learns a dynamic graph prototypes metric space in few-shot HOI.

2) Comparison on NF Split Strategy: The results on the NF datasets are summarized in Table IV. Our DGIG-Net achieves the accuracies of 47.08% on 5-way 1-shot and 73.06% on 5-way 5-shot on HICO-NF, which outperforms the state-of-the-art approaches at least in 3.7% and 2.2%. Similar results are also observed on TUHOI-NF. It can also be observed that the performance on HICO-NF and TUHOI-NF is better than that on HICO-NN and TUHOI-NN. We consider the reason is as follows. From the perspective of data structure, the NN

TABLE V
ABLATION STUDIES OF DGIG-NET ON HICO-NN

Methods	5-way 1-shot	5-way 5-shot
PN	32.56 ± 1.59%	42.49 ± 1.75%
Graph PN	35.69 ± 1.56%	52.34 ± 1.81%
Graph PN + Actions	36.49 ± 1.67%	54.38 ± 1.79%
Graph PN + Actions + R_A	36.86 ± 1.63%	54.89 ± 1.81%
Graph PN + Objects	36.62 ± 1.73%	54.77 ± 1.74%
Graph PN + Objects + R_O	37.23 ± 1.65%	55.13 ± 1.86%
Graph PN + Actions + Objects	38.24 ± 1.60%	56.39 ± 1.83%
DGIG-Net w/o KR-Module	39.02 ± 1.66%	58.38 ± 1.84%
DGIG-Net	**39.13 ± 1.68%**	**59.06 ± 1.89%**

split strategy makes the metatrain set and the metatest set both follow the similar long-tail distribution, which brings difficulty to transfer knowledge among imbalanced class distributions. In contrast, the NF split strategy divides the metatrain set as a head distribution and the metatest set as a tail distribution. It provides much more knowledge from instances. Moreover, the objects and actions in the test set in the NF split strategy may also appear in the metatrain set separately, as the purpose is to recognize unseen combinations.

D. Ablation Studies

We conduct ablation studies to evaluate the impacts of each component in our DGIG-Net in Table VI. We first consider the following variants.

PN is prototypical network [17], which is employed as the baseline for DGIG-Net.

Graph PN applies graph prototypes metric space instead of linear prototypes in PN.

Graph PN + Actions adds action label semantic in Graph PN.

Graph PN + Actions + R_A introduces action label semantic and their dynamic relationship obtained by WordNet [48] in Graph PN.

Graph PN + Objects adds object label semantic in Graph PN.

Graph PN + Objects + R_O introduces object label semantic and its dynamic relationship obtained by WordNet [48] in Graph PN.

Graph PN + Actions + Objects adds both action and object label semantic in Graph PN.

DGIG-Net w/o KR-Module applies both object and action label semantic and their dynamic relationship obtained by WordNet [48] in Graph PN.

First, it is observed that Graph PN improves the results at least 3.1% and 9.8%, respectively, on 5-way 1-shot and 5-way 5-shot compared with PN, as shown in Table V. It proves that the graph prototypes method is more effective than the linear prototypes method for the few-shot HOI recognition. We can observe that applying a single type of semantic information, that is, Graph PN + Actions or Graph PN + Objects, brings at least 0.8% and 2.0% performance gains in terms of both settings, respectively. This is a reasonable phenomenon since introducing the auxiliary semantic information helps to learn a discriminative metric space. In contrast, Graph PN + Actions + Objects applies both types of semantic information, which achieves the surprising accuracies of 38.24% in terms of 5-way 1-shot and 56.39% in terms of 5-way 5-shot. The dynamic relationship role is proved by Graph PN + Actions + R_A and Graph PN + Objects + R_O. They slightly improve the corresponding baseline performance by 0.4% and 0.6%. Remarkably, DGIG-Net (w/o KR-Module), just Graph PN + Actions + Objects + R_A + R_O, achieves 39.02% in terms of 5-way 1-shot and 58.38% in terms of 5-way 5-shot on HICO-NN datasets, which marginally improves the Graph PN + Actions + Objects by 0.7% and 2.0%. It also can be observed that the KR-Module is, respectively, capable of bringing 0.1% and 0.6% performance gains on both settings against DGIG-Net (w/o KR-Module). The KR-Module sees no further improvement on 5-way 1-shot due the intraclass information not working on only 1 sample.

E. Cross-Domain Analysis

To prove the transferability of the proposed approaches, we design cross-domain experiments that are conducted between two datasets with the same split strategy. There are four types of cross-domain settings: 1) HICO-NN→TUHOI-NN; 2) TUHOI-NN→HICO-NN; 3) HICO-NF→TUHOI-NF; and 4) TUHOI-NF→HICO-NF. For the cross-domain setting A→B, the metatrain set of *A* is utilized in the training stage, while metavalidation and metatest of *B* are utilized for validation and evaluation. This cross-domain setting explores whether metalearning could be implemented on data from different source and target domains.

We choose the same comparison algorithm listed in Table III. The results of cross-domain experiments on the NN setting are shown in Table VI. Our DGIG-Net achieves competitive performance, which, respectively, obtains the accuracies of 39.09% and 58.17% on HICO-NN→TUHOI-NN, 38.53% and 56.86% on TUHOI-NN→HICO-NN on 5-way 1-shot and 5-way 5-shot. DGIG-Net performs superior to the second-best approach SGAP-Net, which brings about at least 1.0% and 1.6% performance gains on both 5-way 1-shot and 5-way 5-shot tasks on HICO-NN→TUHOI-NN. However, our DGIG-Net does not show any significant performance improvement in terms of 5-way 5-shot on TUHOI-NN→HICO-NN. It can be further observed that the accuracies of HICO-NN on cross-domain settings are a bit lower than those on the single-domain experiments. We suppose there are fewer instances in the metatrain set of TUHOI-NN than those of HICO-NN, thus TUHOI-NN cannot provide enough transferable knowledge. However, the performance on HICO-NN→TUHOI-NN is superior to that on TUHOI-NN, except for the result of DGIG-Net on 5-way 5-shot. It suggests that the source domain with more samples and categories has a higher adaptation ability on the target domain. Our DGIG-Net achieves incremental improvement, which demonstrates its strong adaptable ability.

We also conduct experiments on the NF setting, which are shown in Table VII. Our proposed approaches achieve at least 4.6% and 2.2% gains on HICO-NF→TUHOI-NF, and 2.2% and 1.8% gains on TUHOI-NF→HICO-NF compared with those of the second-best SGAP-Net [5]. For the comparison between the cross-domain and single-domain experiments,

TABLE VI
CROSS-DOMAIN FEW-SHOT CLASSIFICATION ACCURACY OF DGIG-NET ON HICO-NN→TUHOI-NN AND TUHOI-NN→HICO-NN WITH ± 95% CONFIDENCE INTERVALS

Method	Type	HICO-NN→TUHOI-NN		TUHOI-NN→HICO-NN	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Matching Networks [7]	Metric	29.13 ± 1.48%	38.83 ± 1.78%	31.53 ± 1.62%	38.13 ± 1.67%
Prototypical Networks [17]	Metric	28.20 ± 1.55%	38.93 ± 1.70%	30.13 ± 1.61%	38.97 ± 1.75%
Relation Networks [18]	Metric	28.60 ± 1.69%	35.27 ± 1.72%	30.97 ± 1.63%	36.10 ± 1.73%
DN4 [19]	Metric	35.05 ± 0.93%	46.82 ± 1.79%	28.51 ± 0.83%	37.81 ± 1.70%
TPN [20]	Metric	35.47 ± 1.55%	43.35 ± 1.86%	31.17 ± 1.48%	38.26 ± 1.78%
EGNN [46]	Metric	36.54 ± 1.60%	46.90 ± 1.76%	32.29 ± 1.64%	43.07 ± 1.82%
SGAP-Net [5]	Metric	37.96 ± 1.68%	56.45 ± 1.77%	36.89 ± 1.74%	56.35 ± 1.79%
MAML [22]	Optimization	36.43 ± 1.69%	47.23 ± 1.87%	32.87 ± 1.65%	39.30 ± 1.80%
Reptile [23]	Optimization	37.54 ± 1.73%	46.39 ± 1.85%	33.21 ± 1.51%	39.67 ± 1.79%
DGIG-Net (Ours)	Metric	**39.09 ± 1.52%**	**58.17 ± 1.87%**	**38.53 ± 1.68%**	**56.86 ± 1.76%**

TABLE VII
CROSS-DOMAIN FEW-SHOT CLASSIFICATION ACCURACY OF DGIG-NET ON HICO-NF→TUHOI-NF AND TUHOI-NF→HICO-NF WITH ± 95% CONFIDENCE INTERVALS

Method	Type	HICO-NF→TUHOI-NF		TUHOI-NF→HICO-NF	
		5-way 1-shot	5-way 5-shot	5-way 1-shot	5-way 5-shot
Matching Networks [7]	Metric	36.32 ± 1.64%	53.55 ± 1.72%	33.13 ± 1.57%	47.26 ± 1.70%
Prototypical Networks [17]	Metric	35.05 ± 1.59%	51.26 ± 1.74%	30.30 ± 1.56%	47.70 ± 1.68%
Relation Networks [18]	Metric	41.21 ± 1.73%	53.78 ± 1.85%	36.85 ± 1.61%	47.71 ± 1.77%
DN4 [19]	Metric	38.05 ± 1.69%	55.67 ± 1.98%	32.56 ± 1.45%	49.81 ± 1.73%
TPN [20]	Metric	39.47 ± 1.55%	53.35 ± 1.86%	36.57 ± 1.73%	49.99 ± 1.98%
EGNN [46]	Metric	40.32 ± 1.57%	57.03 ± 1.72%	38.47 ± 1.69%	53.27 ± 1.76%
SGAP-Net [5]	Metric	42.17 ± 1.63%	70.49 ± 1.92%	43.29 ± 1.59%	69.68 ± 1.87%
MAML [22]	Optimization	41.28 ± 1.34%	56.34 ± 1.56%	37.88 ± 1.39%	54.29 ± 1.84%
Reptile [23]	Optimization	41.79 ± 1.77%	57.39 ± 1.81%	38.97 ± 1.52%	54.69 ± 1.78%
DGIG-Net (Ours)	Metric	**46.86 ± 1.56%**	**72.69 ± 1.69%**	**45.47 ± 1.65%**	**71.56 ± 1.74%**

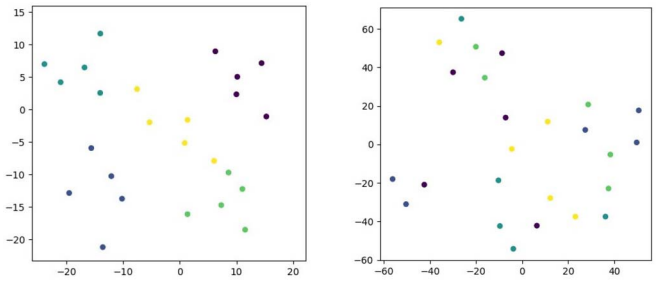
TABLE VIII
COMPLEXITY COMPARISON OF OUR DGIG-NET WITH OTHER METHODS

Methods	Params (M)	FLOPs (G)	Training Time (h)
DN4 [19]	2.472	3.385	29.75
TPN [20]	2.615	3.697	30.53
DGIG-Net (Ours)	3.106	4.294	33.76

the accuracy of TUHOI-NF→HICO-NF achieves 71.56% on 5-way 5-shot, which is 1.5% inferior to that of HICO-NF on the single-domain experiments. Obviously, it can be observed that the results on HICO-NF are better than that on TUHOI-NF→HICO-NF. In contrast, the performance of DGIG-Net on HICO-NF→TUHOI-NF decreases a little, but the other results are better than that of single-domain experiments on TUHOI-NF. From our previous analysis, it depends on the data distribution and data scale of different domains. The cross-domain experiments on both NN and NF settings show that the proposed approaches have robust-domain adaptation ability against the other approaches.

F. Complexity Analysis

We implement the complexity analysis and comparison on our DGIG-Net. For fair comparison, we choose two latest metric-based methods, that is, DN4 [19] and TPN [20]. The total network parameters, floating-point operations, and training time of 500 000 episodes on 5-way 5-shot are listed in Table VIII. The total parameters in our DGIG-Net are 3.106M, and Flops is 4.294G. As for training time, our DGIG-Net spends 33.76 h. It takes about 3 h more than other approaches; however, our DGIG-Net beats them in a large margin on performance.



(a) (b)

Fig. 7. Classification results of few-shot HOI on the HICO-NN dataset. The task setting is 5-way 1-shot. (Best viewed in color.) (a) t-SNE visualization results of DGIG-Net. (b) t-SNE visualization results of prototypical networks.

G. Qualitative Results

1) t-SNE Results: Fig. 8 shows t-SNE [50] visualizations of feature space for our DGIG-Net and prototypical networks [17]. We can observe that our method produces a special discriminative feature space, where class distributions are more distinguishable. Especially, our work applies

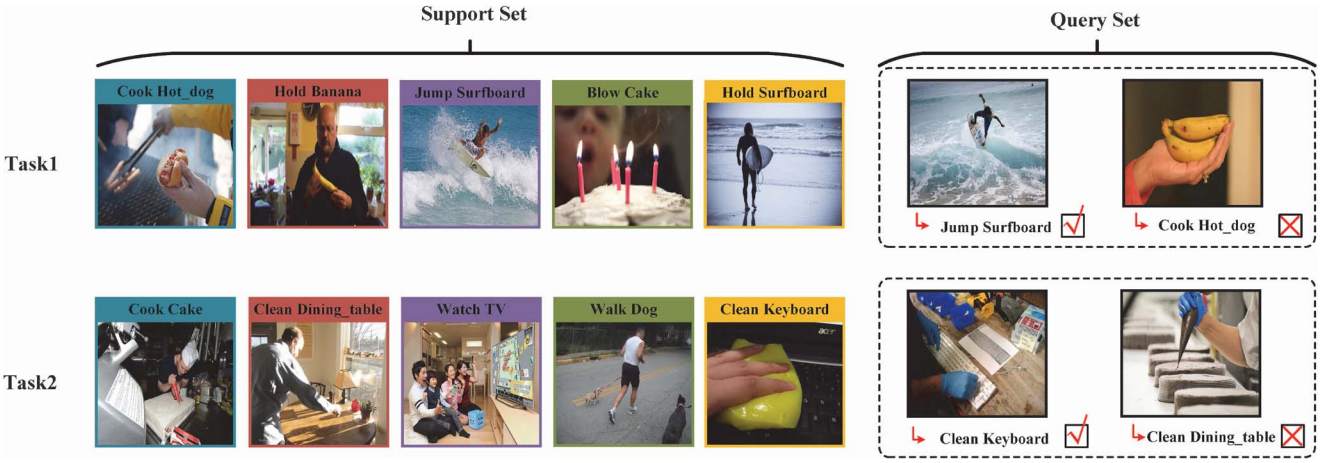


Fig. 8. t-SNE visualization of feature space in terms of 5-way 5-shot on HICO-NN. (a) t-SNE visualization results of DGIG-Net. (b) t-SNE visualization results of prototypical networks.

task-oriented semantic guidance to capture class discriminative information, which learns a dynamic graph-metric space in few-shot HOI. In contrast, the features in prototypical networks are difficult to distinguish [shown in Fig. 8(b)]. This demonstrates that our approach learns a better encoder to represent the features.

2) Classification Results: To further qualitatively verify the effectiveness of our proposed model, we select several representative tasks to show their corresponding few-shot results on HICO-NN. Fig. 7 presents the qualitative results on 5-way 1-shot with two query samples. It can be observed that our model recognizes HOI categories correctly when appearing the same actions or objects in the task. For example, in Task 1, our model can recognize the query image is "Jump-Surfboard" without the interfere of "Hold-Surfboard." On the other side, we are trying to explore the reason for misclassification. Concretely, our model performs unsatisfactorily in: 1) support and query samples present totally different visual angles, such as the two samples of "Hold-Banana" in Task 1. It exists a huge visual bias between local and global views and 2) objects of different shapes. The "Cake" in Task 2 brings coarse-grained and fine-grained level shapes, which requires more detailed information.

V. Conclusion

FSL for HOI is a challenging vision task, in which diversity and interactivity of human actions result in great difficulty to learn adaptive class prototypes. In this work, we have proposed a novel graph prototypes framework, namely, DGIG-Net, to learn a dynamic graph-metric space guided by a task-oriented semantic for few-shot HOI. The KR-Module encodes both the graph structure and node features with a GCN, where the decoder reconstructs the topological graph information and manipulates the latent graph representation. The DR-Module implements a graph-metric space with dynamic task-oriented semantic information to obtain HOI class prototypes. Extensive experiments on four few-shot HOI datasets, HICO-NN, TUHOI-NN, HICO-NF, and TUHOI-NF, have demonstrated that our proposed approaches are superior to state-of-the-art approaches.

References

[1] L. Wu, Y. Wang, X. Li, and J. Gao, "Deep attention-based spatially recursive networks for fine-grained visual recognition," *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1791–1802, May 2019.

[2] Z. Zhang, J. Chen, Q. Wu, and L. Shao, "GII representation-based cross-view gait recognition by discriminative projection with list-wise constraints," *IEEE Trans. Cybern.*, vol. 48, no. 10, pp. 2935–2947, Oct. 2018.

[3] X. Li, M. Chen, F. Nie, and Q. Wang, "A multiview-based parameter free framework for group detection," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4147–4153.

[4] X. Li, M. Chen, F. Nie, and Q. Wang, "Locality adaptive discriminant analysis," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2201–2207.

[5] Z. Ji, X. Liu, Y. Pang, and X. Li, "SGAP-Net: Semantic-guided attentive prototypes network for few-shot human–object interaction recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11085–11092.

[6] Y. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human–object interactions in images," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 1017–1025.

[7] O. Vinyals, C. Blundell, T. P. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3637–3645.

[8] C. Xing, N. Rostamzadeh, B. N. Oreshkin, and P. H. O. Pinheiro, "Adaptive cross-modal few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4848–4858.

[9] P. Tokmakov, Y. Wang, and M. Hebert, "Learning compositional representations for few-shot recognition," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6372–6381.

[10] J. Mu, P. Liang, and N. D. Goodman, "Shaping visual representations with language for few-shot classification," 2019. [Online]. Available: arXiv:1911.02683.

[11] Z. Ji, Y. Sun, Y. Yu, Y. Pang, and J. Han, "Attribute-guided network for cross-modal zero-shot hashing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 321–330, Jan. 2020.

[12] L. Liu, H. Zhang, X. Xu, Z. Zhang, and S. Yan, "Collocating clothes with generative adversarial networks cosupervised by categories and attributes: A multidiscriminator framework," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3540–3554, Sep. 2020.

[13] Z. Ji, J. Yan, Q. Wang, Y. Pang, and X. Li, "Triple discriminator generative adversarial network for zero-shot image classification," *Sci. China Inf. Sci.*, to be published.

[14] H. Li, W. Dong, X. Mei, C. Ma, F. Huang, and B. Hu, "LGM-Net: Learning to generate matching networks for few shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3825–3834.

[15] J. Zhang, M. Zhang, Z. Lu, T. Xiang, and J. Wen, "AdarGCN: Adaptive aggregation GCN for few-shot learning," 2020. [Online]. Available: arXiv:2002.12641.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

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

[18] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[19] W. Li, L. Wang, J. Xu, J. Huo, Y. Gao, and J. Luo, "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7260–7268.

[20] 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.

[21] D. Das and C. S. G. Lee, "A two-stage approach to few-shot learning for image recognition," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 3336–3350, Dec. 2019.

[22] 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.

[23] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018. [Online]. Available: arXiv: 1803.02999.

[24] A. A. Rusu *et al.*, "Meta-learning with latent embedding optimization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–13.

[25] Y. Wang, R. B. Girshick, M. Hebert, and B. Hariharan, "Low-shot learning from imaginary data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7278–7286.

[26] H. Zhang, J. Zhang, and P. Koniusz, "Few-shot learning via saliency-guided hallucination of samples," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2770–2779.

[27] P. Bach, G. Knoblich, T. C. Gunter, A. D. Friederici, and W. Prinz, "Action comprehension: Deriving spatial and functional relations," *J. Exp. Psychol. Human Perception Perform.*, vol. 31, no. 3, pp. 465–479, 2005.

[28] A. Gupta and L. S. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[29] D. Le, J. R. R. Uijlings, and R. Bernardi, "TUHOI: Trento universal human object interaction dataset," in *Proc. Workshop Vis. Lang.*, 2014, pp. 17–24.

[30] G. Gkioxari, R. Girshick, and J. Malik, "Actions and attributes from wholes and parts," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 2470–2478.

[31] H. Fang, J. Cao, Y. Tai, and C. Lu, "Pairwise body-part attention for recognizing human–object interactions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 52–68.

[32] A. Mallya and S. Lazebnik, "Learning models for actions and person-object interactions with transfer to question answering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 414–428.

[33] G. Gkioxari, R. B. Girshick, P. Dollar, and K. He, "Detecting and recognizing human–object interactions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8359–8367.

[34] Y. Li *et al.*, "Transferable interactiveness knowledge for human–object interaction detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3585–3594.

[35] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human–object interaction," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–251.

[36] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," 2019. [Online]. Available: arXiv: 1901.00596.

[37] S. Qi, W. Wang, B. Jia, J. Shen, and S. C. Zhu, "Learning human–object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 407–423.

[38] M. D. G. Mallea, P. Meltzer, and P. J. Bentley, "Capsule neural networks for graph classification using explicit tensorial graph representations," 2019. [Online]. Available: arXiv:1902.08399.

[39] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[40] R. Li, S. Wang, F. Zhu, and J. Huang, "Adaptive graph convolutional neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3546–3553.

[41] C. Zhuang and Q. Ma, "Dual graph convolutional networks for graph-based semi-supervised classification," in *Proc. Web Conf.*, 2018, pp. 499–508.

[42] X. Zhou *et al.*, "Graph convolutional network hashing," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1460–1472, Apr. 2020.

[43] V. G. Satorras and J. B. Estrach, "Few-shot learning with graph neural networks," in *Proc. ICLR*, 2018, pp. 1–13.

[44] S. Gidaris and N. Komodakis, "Generating classification weights with gnn denoising autoencoders for few-shot learning," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 21–30.

[45] A. Iscen, G. Tolias, Y. Avrithis, O. Chum, and C. Schmid, "Graph convolutional networks for learning with few clean and many noisy labels," 2019. [Online]. Available: arXiv:1910.00324.

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

[47] C. Wang, S. Pan, R. Hu, G. Long, J. Jiang, and C. Zhang, "Attributed graph clustering: A deep attentional embedding approach," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3670–3676.

[48] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

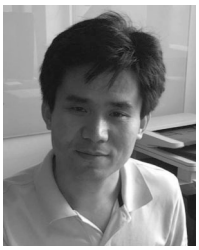
[50] L. V. D. Maaten and G. E. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

Xiyao Liu (Graduate Student Member, IEEE) received the B.S. degree in telecommunication engineering from Tianjin University, Tianjin, China, in 2015, where she is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering.

Her research interests include few-shot learning, human–object interaction, and computer vision.

Zhong Ji (Senior Member, IEEE) received the Ph.D. degree in signal and information processing from Tianjin University, Tianjin, China, in 2008.

He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University. He has authored more than 80 technical articles in refereed journals and proceedings, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, PR, CVPR, ICCV, ECCV, NeurIPS, AAAI, and IJCAI. His current research interests include multimedia understanding, zero/few-shot learning, cross-modal analysis, and video summarization.



Yanwei Pang (Senior Member, IEEE) received the Ph.D. degree in electronic engineering from the University of Science and Technology of China, Hefei, China, in 2004.

He is currently a Professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. He has authored over 120 scientific papers. His current research interests include object detection and recognition, vision in bad weather, and computer vision.

Jungong Han (Senior Member, IEEE) received the Ph.D. degree in telecommunication and information system from Xidian University, Shaanxi, China, in 2004.

He is currently a Full Professor and the Chair of Computer Science with Aberystwyth University, Aberystwyth, U.K. He has published over 180 papers, including over 50 IEEE TRANSACTIONS and over 40 A* conference papers. His research interests span the fields of video analysis, computer vision, and applied machine learning.

Xuelong Li (Fellow, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2002.

He is a Full Professor with School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China.