Semantic-visual shared knowledge graph for zero-shot learning

Beibei Yu [1] , **Cheng Xie** [Corresp., 1] , **Peng Tang** [1] , **Bin Li** [1]

[1] School of Software, Yunnan University, Kunming, Yunnan, Chain

Corresponding Author: Cheng Xie
Email address: xiecheng@ynu.edu.cn

Almost existing zero-shot learning methods work only on benchmark datasets (E.g., CUB, SUN, AwA, FLO and aPY) which have already provided pre-defined attributes for all the classes. These methods thus are hard to apply on real-world datasets (like ImageNet) since there are no such pre-defined attributes in the data environment. Latest works have explored to use semantic-rich knowledge graphs (such as WordNet) to substitute pre-defined attributes. However, these methods encounter a serious " role="presentation"> domain shift'' problem because such knowledge graph cannot provide detail enough semantics for describing fine-grained information. To this end, we propose a semantic-visual shared knowledge graph (SVKG) to enhance the detailed information for zero-shot learning. SVKG represents high-level information by using semantic embedding but describes fine-grained information by using visual features. These visual features can be directly extracted from real-world images to substitute pre-defined attributes. A multi-modals graph convolution network is also proposed to transfer SVKG into graph representations that can be used for downstream zero-shot learning tasks. Experimental results on the real-world datasets without pre-defined attributes demonstrate the effectiveness of our method and show the benefits of the proposed. And our method obtains a +2.8%, +0.5%, and +0.2% increase compared with the state-of-the-art in 2-hops, 3-hops, and all divisions relatively.

Semantic-Visual Shared Knowledge Graph for Zero-Shot Learning

Beibei Yu[1], Cheng Xie[1], Peng Tan[1], and Bin Li[1]

[1]School of Software, Yunnan University, Kunming, China

Corresponding author:
Cheng Xie[1]

Email address: xiecheng@ynu.edu.cn

ABSTRACT

Almost existing zero-shot learning methods work only on benchmark datasets (E.g., CUB, SUN, AwA, FLO and aPY) which have already provided pre-defined attributes for all the classes. These methods thus are hard to apply on real-world datasets (like ImageNet) since there are no such pre-defined attributes in the data environment. Latest works have explored to use semantic-rich knowledge graphs (such as WordNet) to substitute pre-defined attributes. However, these methods encounter a serious "domain shift" problem because such knowledge graph cannot provide detail enough semantics for describing fine-grained information. To this end, we propose a semantic-visual shared knowledge graph ($SVKG$) to enhance the detailed information for zero-shot learning. $SVKG$ represents high-level information by using semantic embedding but describes fine-grained information by using visual features. These visual features can be directly extracted from real-world images to substitute pre-defined attributes. A multi-modals graph convolution network is also proposed to transfer $SVKG$ into graph representations that can be used for downstream zero-shot learning tasks. Experimental results on the real-world datasets without pre-defined attributes demonstrate the effectiveness of our method and show the benefits of the proposed $SVKG$. And our method obtains a +2.8%, +0.5%, and +0.2% increase compared with the state-of-the-art in 2-hops, 3-hops, and all divisions relatively.

INTRODUCTION

In recent years, zero-shot learning has attracted widespread attention in computer vision and machine learning areas. It aims to predict the new classes that have never been appeared during the training process. The base idea of zero-shot learning is to use labeled semantic information (normally word attributes) to learn a projection between semantic space and visual space. Then, visual samples from new classes can be projected into semantic space to match the attributes to decide the corresponding classifications. In the past five years, a large number of methods for zero-shot learning have been proposed based on this idea (Liu et al., 2022; Kim et al., 2021; Chen et al., 2021, 2020b; Annadani and Biswas, 2018; Li et al., 2020; Yu and Lee, 2019; Liu et al., 2019; Cacheux et al., 2019).

Table 1. Six traditional benchmark datasets with pre-defined attributes, for zero-shot learning only.

	Seen classes	Seen samples	Unseen classes	Unseen samples	**Attributes**
aPY	20	15,399	12	7,924	**64**
AwA1	40	30,475	10	5,685	**85**
AwA2	40	37,322	10	7,913	**85**
FLO	82	8,189	20	1,155	**1,024**
CUB	150	11,788	50	2,967	**1,024**
SUN	645	14,340	72	1,440	**102**
Comparing with real-world dataset					
ImageNet	1K	1.29M	20K	12.9M	**0**

However, these zero-shot learning methods inevitably need to follow a major premise that semantic attributes should be pre-defined and labeled in the datasets. Almost all recent zero-shot learning works are only evaluated on six small benchmark datasets (CUB, SUN, aPY, FLO, AwA1 and AwA2). Table 1 provides the statistics of these datasets. It can be observed the semantic attributes are pre-defined in each dataset from 64 (aPY) to 1024 (CUB) which describe fine-grained semantics for all classes like "Wings-Color", "Leg-color", "Breast", etc. Moreover, these datasets are quite small, only tens or hundreds classes and tens of thousands samples, which are far away from real-world data environment, normally tens of thousands classes and millions samples. Thus, existing zero-shot methods are hard to be applied in real-world data environments such as ImageNet which does not provide any pre-defined attributes for any classes.
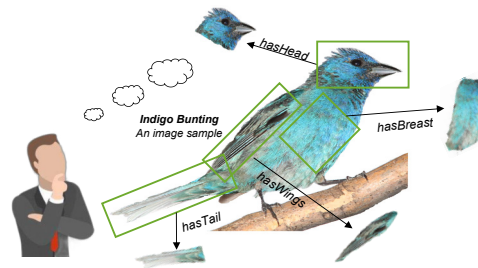
Figure 1. Fine-grained semantics can be extracted from a raw image sample. It can be further used to establish a fine-grained knowledge graph through extracting the amount of image samples.

In the past four years, the graph neural network (GNN) has been adopted by zero-shot learning, which makes zero-shot learning tasks applicable in real-world data environments (Wang et al., 2018a; Kampffmeyer et al., 2019a; Wang and Jiang, 2021a). This is because GNN-based methods use open knowledge (Wikipedia, freebase, Nell, etc.) to substitute pre-defined attributes in semantics and have better generalization power in semantic representation. However, these GCN-based zero-shot learning methods for image classification on the ImageNet dataset, like GCNZ (Wang et al., 2018a),DGP(Kampffmeyer et al., 2019a) and CL-DKG(Wang and Jiang, 2021a), still have a long way from the traditional methods evaluated in the six small benchmark datasets, which normally achieve more than 50% Top-1 accuracy. While the out-of-state method(Wang and Jiang, 2021a) only obtains 1.0% Top-1 accuracy and 16.7% Top-20 accuracy on ImageNet dataset. The main finding of our analysis is that, in comparison to the pre-defined attributes in the six benchmark datasets, the knowledge network used in GCN-based approaches offers significantly less fine-grained semantics. This leads a serious "domain-shift" problem (Min et al., 2020; Fu et al., 2015; Zhu et al., 2018; Wan et al., 2019; Ni et al., 2019) during projection between semantic space and visual space. In other words, an existing knowledge graph can only provide coarse-grained semantics (such as name, habitat, taxonomy, hypernym, hyponym, etc.) without fine-grained semantics (like head, tail, beak, legs, etc.). Thus, the core question becomes "can a knowledge graph provide fine-grained semantics or even more for zero-shot learning? ".

Based on the above discussion, we are going to present a solution that lets fine-grained semantics be extracted from raw images and shared with the existing knowledge graph. The idea is that there are lots of fine-grained semantics already hidden in the image samples, especially in the large-scale image set. As the example shown in Figure 1., when a person sees an image sample about "Indigo Bunting", he can quickly establish fine-grained semantics about this bird in his brain, such as "has deep and light blue head", "has black and light blue tail", "has deep blue breast". Then, a large fine-grained semantic network (graph) can be further established if there are enough image samples have been seen. On the contrary, these fine-grained semantics can also complement semantic representations.

Based on the idea, we first scan all the image samples of seen classes in ImageNet-1K to extract parts of visual features for each seen class. Then, WordNet nodes and relations are extracted according to the seen and unseen class labels in ImageNet-1K and embedded as semantic features by a word embedding model. After that, parts of visual features and WordNet semantic features are connected together as a semantic-visual shared knowledge graph ($SVKG$). At last, a multi-modals GCN network is proposed to embed ($SVKG$) into graph representations that can be used in zero-shot learning tasks, as showed in Figure 2. Experimental results on real-world datasets without pre-defined attributes demonstrate the

effectiveness of our method and show the benefits of the proposed semantic-visual shared knowledge graph.

Overall, the main contributions of this paper are summarized as follows:

- We first propose the semantic-visual shared knowledge graph that stores high-level information in semantic features while storing fine-grained information in parts visual features. It can be used for real-world zero-shot learning tasks without requiring pre-defined attributes.

- We propose a multi-modal GCN network to fuse semantic modal and visual modal together in the same knowledge graph. The output of the network is semantic-visual shared graph representation that can be easily used for downstream zero-shot learning tasks.

RELATED WORK

Zero-Shot Learning based on attributes

Attribute-based methods account for the largest proportion of zero-shot learning (ZSL) research since ZSL was first proposed in DAP (Lampert et al., 2009) in 2009 with a proposed dataset called AwA which contains pre-defined attributes. Its principle is to annotate the attributes of images (such as whether there is a tail, hair color, etc.), then learn the semantic attribute characteristics of visual objects, and finally judge whether the attribute combinations are satisfied by the visual objects.

Following the design principles of AwA dataset, a series of benchmark datasets with pre-defined attributes are established today, including AwA2, CUB, SUN, FLO and aPY. Based on these benchmark datasets, ZSL has a blowout development with several milestones.

In 2013, with the development of semantic embedding technology, the first milestone of ZSL was the ALE model (Akata et al., 2013) proposed by Akata et al., which can encode the attributes as semantic vectors, encode images as visual vectors, and then learn a function to calculate the similarity between semantic vectors and visual vectors, so as to match the corresponding mappings between attributes and images.

In 2017, with the development of deep learning technology in the field of visual computing, the second milestone of ZSL is the SAE model (Kodirov et al., 2017) proposed by kodirov et al. By using the Auto-Encoder network, it can encode more fine-grained attribute features and visual features, so as to better achieve "semantic-to-visual" matching. The overall performance of SAE is significantly improved compared with ALE.

In 2018, with the remarkable performance of Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) in image processing, ZSL achieved the third milestone with the representative model GAZSL (Zhu et al., 2018) proposed by Y. Zhu et al. It uses GAN to synthesize fake visual features from semantic features, and then match the fake visual features with the real visual features to predict the unseen objects.

Recently, some researchers have started to investigate how to mix various types of semantics (attributes, word2vec, text description, knowledge graph) to address further "domain shift" issues. For instance, (Wang et al., 2021) learns discriminative classifiers using a variety of semantic viewpoints. In (Xie et al., 2021; Naeem et al., 2021), open knowledge is regarded as auxiliary or augmented semantics added with pre-defined features.

However, the above methods use pre-defined properties as their primary semantic source. As a result, these methods are essentially restricted to benchmark datasets with pre-defined features. The applicability of these methods in real-world environments devoid of pre-defined features is restricted to a small number of cases.

Zero-Shot Learning based on knowledge graph

Knowledge graph (KG) actually is a third part knowledge base that can provide semantic information for semantic-to-visual transformation in zero-shot learning. Knowledge graph (KG) actually constitutes a third-party knowledge base that can provide semantic information for semantic-to-visual transformation in zero-shot learning. Thus, KG is intuitively considered a substitution for pre-defined attributes. Benefiting from the development of GNN, GCNZ(Wang et al., 2018a) creates a formal knowledge graph based on WordNet to substitute pre-defined attributes and learn semantic embedding from structure information and word embedding. It is, indeed, the first work that tries to apply knowledge graphs and GNN as the backbone for the real-world dataset zero-shot learning task. Based on (Wang et al., 2018a), (Kampffmeyer
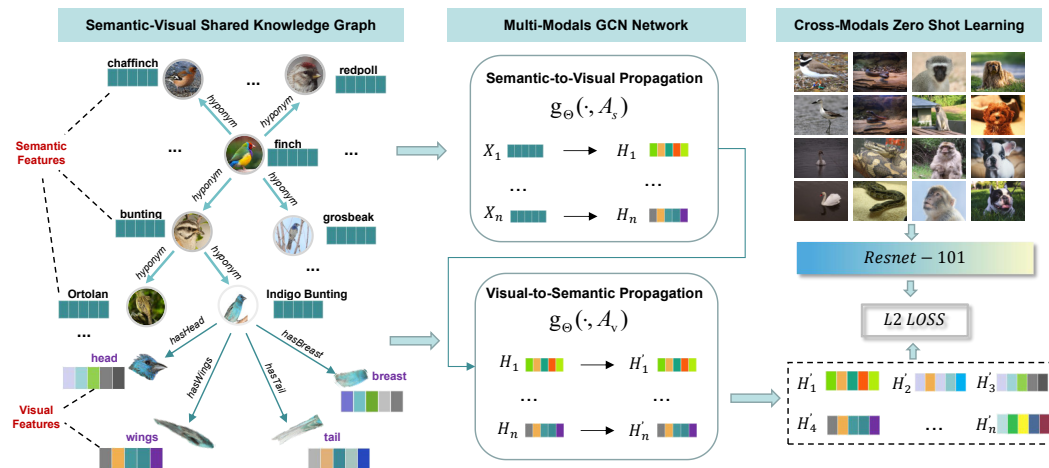
Figure 2. The overview of the proposed method. It can be divided into three modules. The first module, semantic-visual shared knowledge graph, is introduced in section "Semantic-visual Shared Knowledge Graph"; The second module, multi-modal GCN network, is detailed in section "Multi-modals Graph Representation Learning"; The last module, zero-shot learning, is explained in section "Cross-modals Zero-shot Learning"

126   et al., 2019a) proposes a so-called Dense Graph Propagation (DGP) method to gather the semantic
127   information through the relations of the knowledge graph. And, CL-DKG (Wang and Jiang, 2021a)
128   applies contrastive learning on dual knowledge graph to learn the projection between semantic space
129   and visual space without any pre-defined attributes. It exploits multiple knowledge relationships among
130   classes simultaneously to learn robust and discriminative classifiers for unseen classes.
131   However, compared to the pre-defined features in the six benchmark datasets, the knowledge network
132   used in GNN-based approaches (Wang et al., 2018a; Kampffmeyer et al., 2019a; Liu et al., 2020; Wang
133   et al., 2021) provides far less fine-grained semantics. In contrast to them, we propose a solution that
134   enables fine-grained semantics to be extracted from raw images and shared with the existing knowledge
135   graph in order to enhance recognition performance in the ZSL challenge.

Visual knowledge applied for zero-shot learning

137   To enhance the fine-grained semantic information for existing KG, the latest research tries to apply "visual
138   knowledge" in zero-shot learning. Zhu et al. (2019) propose a novel low-dimensional embedding for visual
139   objects called "visually semantic" to narrow the semantic gap between high-dimensional visual space
140   and semantic space in zero-shot learning. Xu et al. (2020) proposed a zero-shot learning framework that
141   jointly learns discriminative visual semantics only using class-level attributes. Liu et al. (2021) propose a
142   goal-oriented gaze estimation module (GEM) to improve the discriminative attribute localization based
143   on the class-level attributes for ZSL. They introduce so-called "human gaze locations" to obtain new
144   regions of visual semantics. Xie et al. (2020) extracts the relationships among visual regions to enhance
145   the semantic information for both seen and unseen classes and then can lunch region-based relational
146   reasoning for ZSL. Song and Zhang (2022) consists of two graphs constructed by semantic and visual
147   representations respectively to enhance semantic representations.
148   However, the above methods continue to see visual semantics as auxiliary or augmented semantics.
149   In contrast to them, we want to combine semantic and visual representations into a single common
150   knowledge graph, so-called visual-semantic shared knowledge graph.

METHODOLOGY

Problem definition.

153   Knowledge Graph based Zero-Shot learning uses the knowledge-aided method to learn an image classifica-
154   tion model from seen classes to predict unseen classes. First, given a knowledge graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, Then,
155   we have $X \in \mathbb{R}^{N \times F}$ be the word-embedding set for all the nodes in $\mathscr{V}$, and $A \in \mathbb{R}^{N \times N}$ be the adjacency

156 matrix transferred from $\mathscr{E}$. Here, F is the dimensions of the embedding. After, a graph representation
157 model $g_\Theta(X, A, I_{seen})$ is learned to transfer X to the graph node representation $H \in \mathbb{R}^{N \times F'}$ supervised by
158 I_{seen}. Here, $I_{seen} \in \mathbb{R}^{N \times F''}$ is the seen image feature set extracted by a CNN model. At last, an image
159 classification model $\mathscr{L}_\Theta(H, I_{unseen})$ is learned to classify I_{unseen} to the particular class according to the
160 graph node representation H. Here, $I_{unseen} \in \mathbb{R}^{N \times F''}$ is the unseen image feature set extracted by a CNN
161 model.

Semantic-visual Shared Knowledge Graph

163 Semantic-Visual Shared Knowledge Graph ($SVKG$) is a multi-modal graph that contains both semantic
164 embedding and visual embedding in the same graph. Let X represents the embedding set of the graph
165 nodes with X_s denotes Word-Embedding nodes and X_v represents CNN-Embedding nodes. A represents
166 the edge set while A_s denotes the edges among word-Embedding nodes and A_v denotes the edges among
167 Resnet-Embedding nodes. $SVKG$ is defined in Algorithm.(1).

Algorithm 1 Semantic-visual shared knowledge graph

Input: $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, seen images
Output: $SVKG$
1: $X_s \longleftarrow$ Put X into Glove
2: $A_s \longleftarrow$ Put X into WordNet hyponym/hypernym
3: $X_v \longleftarrow$ Put seen images into EfficientDet
4: $A_s \longleftarrow$ Put X_v connecte with semantic object node
5: $X' \longleftarrow X_s \cup X_v$
6: $A' \longleftarrow A_s \cup A_v$
7: $svkg \longleftarrow \{ X', A' \}$
8: **return** $SVKG$

---

168 In $SVKG$, the semantic node set X_s is constructed from WordNet[1] Noun words and embedded into
169 word features by Glove[2]. The edge matrix A_s is established by WordNet hyponym/hypernym links
170 among these words. The visual node set X_v is obtained by using EfficientDet Tan et al. (2020) to detect
171 fine-grained parts visual features (such as "head", "back", "belly", "breast", "leg", etc) from ImageNet-1k.
172 Then, these visual nodes are connected to their semantic object node by the edge matrix A_v (such as
173 "hasHead", "hasBelly", etc). An example of $SVKG$ about bird Finch is presented in Figure 3.
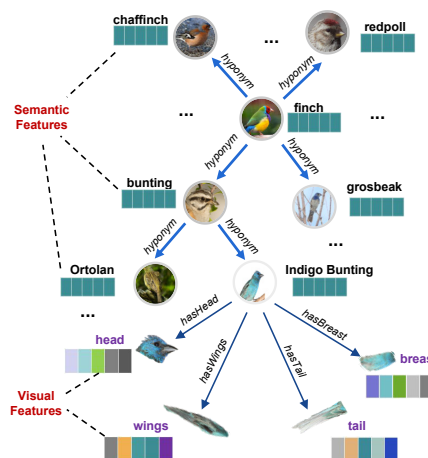
Figure 3. An example of Semantic-visual Shared Knowledge Graph ($SVKG$) about bird Finch

---

[1] https://wordnet.princeton.edu/
[2] https://nlp.stanford.edu/projects/glove/

Graph Augmentation for $SVKG$

In the zero-shot learning process, semantic feature space X_s needs to be transferred into semantic-visual shared space H. However, the visual feature space (fine-grained visual parts) X_v might cause interference during $X_s \longrightarrow H$ transfer process, leading to a serious over-fitting problem. Empirically, many data augmentation methods have shown to improve the generalization and robustness of the learning model Zhao et al. (2021); Rong et al. (2020); Srivastava et al. (2014); Chen et al. (2020a). Thus, in this work, we introduce an extra graph node with zero-initialized features to augment $SVKG$. This zero-initialized node X_{aug} then links to all seen nodes by the edge matrix A_{aug}. After that, the node is indirectly connected to all visual feature nodes X_v through the corresponding seen nodes. It can moderate the strength of feature propagation from visual feature node X_v to alleviate the over-fitting problem. Augmentation graph $SVKG_{aug}$ is defined in Eq.(1).

$$SVKG_{aug} = (X', A')$$
$$X' = X \cup X_{aug}, \; A' = A \cup A_{aug} \tag{1}$$

Where X_{aug} is the augmentation node with zero-initialized feature. A_{aug} is the adjacency matrix to link X_{aug} to all seen nodes. And X', A' represents feature matrix and adjacency matrix of $SVKG$, respectively. An example of $SVKG_{aug}$ is presented in Figure 4.
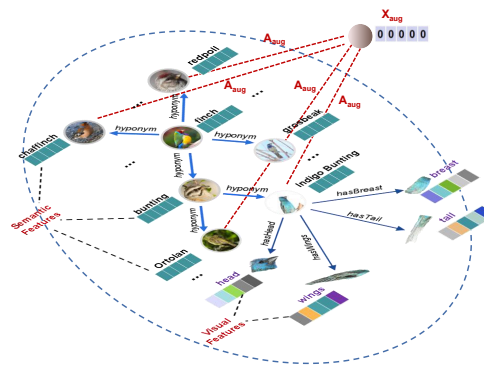
Figure 4. An example of augmentation knowledge graph ($SVKG_{aug}$) about bird Finch.

Multi-modals Graph Representation Learning

As explained in Section. (Problem Definition), knowledge graph $SVKG_{aug}$ needs to be represented in a particular feature space to support the downstream task (zero-shot image classification). In $SVKG_{aug}$, there are two feature modals. One is semantic feature modal $\{X_s$ generated from Glove-Embedding. The other is visual feature modal X_v extracted by EfficientDet. Thus, multi-modals graph representation learning is aimed to transfer both X_s and X_v into a semantic-visual shared feature modal H.

More specifically, we propose a dual ways propagation graph convolution network to transfer X_s and X_v into H, as defined in Eq.(2).

$$X \longrightarrow H \longrightarrow H'$$
$$H = g_\Theta(\sigma(D_s^{-1}(A_s \cup A_{aug})X))$$
$$H' = g_\Theta(\sigma(D_v^{-1}(A_v \cup A_{aug})H)) \tag{2}$$

Where g_Θ is a Graph Convolutional Network (GCN) employed from Kipf and Welling (2017). $\sigma(\cdot)$ represents a nonlinear activation function. D is the diagonal degree matrix of A matrix, where $D_{(i,i)} = \sum_j A_{(i,j)}$. H' is the semantic-visual shared graph representation feature set that can be used for downstream task.

Cross-modals Zero-shot Learning

So far, we already obtained the final representation H' for the knowledge graph $SVKG_{aug}$. Each H'_i represents the centroid feature of the corresponding class (this class could be seen class or unseen class). Here, H' is knowledge graph feature modal. However, in the zero-shot image classification, the targets are the real images which is an image feature modal. Thus, the problem becomes a cross-modal transfer problem (Zhuang et al., 2021). Let I represents the real image features. I_{seen} and I_{unseen} denote seen classes features and unseen classes features relatively. $I_{i,seen}$ represents the centroids feature of ith seen class. Then, we have following the loss function, Eq.(3), to force knowledge graph features H' close to real image features I.

$$H' \longrightarrow differences \longleftarrow I$$

$$\mathscr{L} = \frac{1}{N}\sum_{i=1}^{N}\left(H'_i - I_{i,seen}\right)^2 \tag{3}$$

The base idea is simple. In the training process, the graph features H' could be updated supervised by real image features I. We calculate the similarities for all $(H'_i - I_{i,seen})$ pairs and use average similarity differences as the loss function to train the zero-shot classification model. Note that, in the experiment, a pre-trained Resnet-50 model is used to extract real image features I.

In the predicting process, we first use the same Resnet-50 model to extract unseen image feature I_{unseen} (Algorithm.2 Input). Then, I_{unseen} is input into the trained classification model to calculate the similarities with all graph features H'. To obtain the predicting result, $Result_{set}$ records all the similarities between H' and I_{unseen} ,and the $label_i$ of corresponding $H'_{i]}$ (Algorithm.2 2-6). After, the top-k highest similarities with labels are selected as the zero-shot classification result (Algorithm.2 7-9). Algorithm.2 presents the whole zero-shot predicting process.

Algorithm 2 Zero-shot Predicting Process

Input: H', I_{unseen}
Output: Top-k result
1: $Result_{set} \longleftarrow \{\emptyset, \emptyset\}$
2: **for** H'_i in H' **do**
3:     $Sim_i \longleftarrow H'_i \cdot I_{unseen}$
4:     $lablel_i \longleftarrow$ findLabel(H'_i)
5:     $Result_{set} \longleftarrow \{Sim_i, lablel_i\}$
6: **end for**
7: $Result_{set} \longleftarrow$ SortBySim($Result_{set}$)
8: $Result_{set} \longleftarrow$ Top-k($Result_{set}$)
9: **return** $Result_{set}$

It can be observed, the unseen image feature I_{unseen} is not used during the whole training process. In the predicting process, the input I_{unseen} is the first time the model touches the unseen images.

EXPERIMENTS

In this section, the datasets and evaluation metrics are introduced first. Then, the implementation details about the model settings are explained. After, the state-of-the-arts comparisons are presented. Then, an ablation study is conducted for the proposed $SVKG$. At last, extra downstream tasks and observations are discussed.

Datasets and Evaluation Metrics

The experiments are conducted on ImageNet (Deng et al., 2009), which is a real-world dataset and also the most large-scale benchmark for zero-shot image classification. The divisions for zero-shot tasks are 1K seen classes from ImageNet-2012-1K while 1.5K, 7.8K and 20K unseen classes for 2-hops, 3-hops and all from ImageNet. These divisions are grouped together in our experiments named "General" group. Here, n-hops means the most n^{th} jumps to connect to other classes of ImageNet through the relations of a

WordNet graph. For example, as showed in Figure 2, given a seen class "Indigo Bunting", 2-hops can be unseen classes "Ortolan" and "Finch", but 3-hops will has extra unseen classes including "Chaffinch", "Redpoll" and "Grosbeak". Note that there is no overlap between the seen and unseen classes in the three divisions.

In order to show the effectiveness of visual feature X_v in $SVKG$, a "Detailed" group which contains four subsets ('bird', 'snake', 'primate' and 'dog' categories) are divided from ImageNet-1K and corresponding parts visual features X_v for all these detailed categories are extracted from their image samples by EfficientDet. In a short, the experiments involve two groups and seven divisions divided by ImageNet. The detailed information for each corresponding division is shown in Table 2.

Table 2. The datasets divided from ImageNet

Group	Division	Train		Test		label
		cls	samples	cls	samples	
General	2-hops	1k	1.29M	1.5k	1.3M	38.6%
	3-hops	1k	1.29M	7.8k	5.8M	11.3%
	All	1k	1.29M	20k	12.9M	4.6%
Detailed*	bird	58	74k	635	579k	8.4%
	snake	17	17k	98	76k	14.8%
	primate	19	25k	51	35k	27.1%
	dog	118	155k	77	73k	60.5%

The "Detailed" group can be downloaded from https://drive.google.com/drive/folders/19
Hg59bfIusNLNLONkxoswPbsCUGpb$_G$d?usp = sharing.

There, the label rate demonstrates the difficulty of the corresponding division. This is because with the reduction of supervision in semi-supervised manner, there are higher requirements for the generalization and robustness of the model. The performance of corresponding division will also degrade as the supervised label rate decreases. Therefore, the accuracy in widely different divisions can reflect the robustness of the model. As can be seen from Table 2, the most difficult task is the all division. The label rate of the four subsets increases in turn. Among them, the bird division simulates the division of "All", which most intuitively illustrates the impact of fine-grained semantics. In contrast, the dog division is completely different from the "All" division where the majority of classes are seen.

The proposed method is evaluated according to Generalized Zero-Shot Learning (GZSL) setting, which is the most challenging evaluation metric for zero-shot learning. In GZSL, all classes, no matter seen or unseen classes, are all considered as the candidate classes in the testing, while the test samples are only from unseen classes. We adopt the same train/test splits and the Top-k Hit Ratio (Hit@k) metric, in accordance with (Wang et al., 2018b; Kampffmeyer et al., 2019b; Wang and Jiang, 2021b) for all divisions.

Implementation Details

In the model implementations, Resnet-50 (He et al., 2016) is used as feature extractor to extract 2048-dimensions visual feature that has been pre-trained on the ImageNet 2012 dataset. GloVe (Pennington et al., 2014) is used to extract 300-dimensions semantic embedding for initial graph representation on WordNet nodes. EfficientDet (Tan et al., 2020) is used to extract 300-dimensions visual features for initial graph representation on fine-grained part nodes. The proposed method is trained for 1000 epochs using ADAM (Kingma and Ba, 2015) optimizer with learning rate 0.001 and weight decay 0.0005. For each convolutional layer, we employ Dropout operation and leaky ReLUs with a dropout rate of 0.5 and a negative slope of 0.2 respectively. Besides, the model is implemented by PyTorch, training on 4×GTX-2080Ti GPUs.

Comparisons with State-of-the-Art

In this section, all the comparisons conducted in both General and Detailed groups follow the GZSL setting. For General group, DeViSE (Frome et al., 2013), ConSE (Norouzi et al., 2014), ConSE2 (Wang et al., 2018b), GCNZ (Wang et al., 2018b), DGP (Kampffmeyer et al., 2019b) and CL-DKG (Wang and Jiang, 2021b) are selected as counterparts. The results of these counterparts are directly copied from their

270 original publications. For Detailed group, we apply the sources code of each counterpart to conduct the
271 comparison in four sub-categories datasets. Only DGP, GCNZ and SGCN (Kampffmeyer et al., 2019b)
272 are compared since other methods have not provided the source code.

Table 3. Top-k evaluation in General Group

Test set	Model	hit@k(%)				
		1	2	5	10	20
2-hops(+1k)	DeViSE (Frome et al., 2013)	0.8	2.7	7.9	14.2	22.7
	ConSE (Norouzi et al., 2014)	0.3	6.2	17	24.9	33.5
	ConSE2 (Wang et al., 2018b)	0.1	11.2	24.3	29.1	32.7
	GCNZ (Wang et al., 2018b)	9.7	20.4	42.6	57	68.2
	DGP (Kampffmeyer et al., 2019b)	10.3	26.4	50.3	65.2	76
	CLDGK (Wang and Jiang, 2021b)	7.0	**26.8**	**52.5**	**67.5**	**77.9**
	ours	**13.1**	25.7	47.0	61.0	72.4
3-hops(+1k)	DeViSE (Frome et al., 2013)	0.5	1.4	3.4	5.9	9.7
	ConSE (Norouzi et al., 2014)	0.2	2.2	5.9	9.7	14.3
	ConSE (Wang et al., 2018b)	0.2	3.2	7.3	10	12.2
	GCNZ (Wang et al., 2018b)	2.2	5.1	11.9	18	25.6
	DGP (Kampffmeyer et al., 2019b)	2.9	**7.1**	16.1	24.9	35.1
	CLDGK (Wang and Jiang, 2021b)	2.0	**7.1**	**17.3**	**26.2**	**36.5**
	ours	**3.4**	6.9	14.9	22.7	32.2
All(+1k)	DeViSE (Frome et al., 2013)	0.3	0.8	1.9	3.2	5.3
	ConSE (Norouzi et al., 2014)	0.2	1.2	3	5	7.5
	ConSE (Wang et al., 2018b)	0.1	1.5	3.5	4.9	6.2
	GCNZ (Wang et al., 2018b)	1.0	2.3	5.3	8.1	11.7
	DGP (Kampffmeyer et al., 2019b)	1.4	**3.4**	7.9	12.6	18.7
	CLDGK (Wang and Jiang, 2021b)	1.0	**3.4**	**8.5**	**13.2**	**19.3**
	ours	**1.6**	3.3	7.2	11.3	16.7

273 **General Group Comparisons**: From Table 3, it can be observed that our method achieves a new
274 state-of-the-art on Top-1 accuracy in all divisions of ImageNet. Especially, our method obtains +2.8%,
275 +0.5% and +0.2% increasing compared with DGP in 2-hops, 3-hops and all divisions relatively. It is a
276 significant progress since Top-1 accuracy in ImageNet zero-shot learning is the most challenging and
277 representative task. The result also demonstrates that fine-grained semantic and visual features guide
278 model generates more discriminative graph representations, which are helpful for recognizing fine-grained
279 unseen classes.

280 But from Top-2 to Top-20 accuracy evaluations, the latest method CL-DKG surpasses our method.
281 The reason might be the fine-grained features strengthen discrimination ability but weaken, to some extent,
282 the generalization ability of the model. However, it can be seen from Table 3, the results of our method
283 are very close to CL-DKG and DGP that are still comparable with the state-of-the-art.

284 **Detailed Group Comparisons**: The most significant contribution of the proposed method is that
285 fine-grained visual features are shared with semantic features in the same knowledge graph. Here, Detail
286 comparisons are aimed to show the effects of such shared visual features through some representative
287 categories of ImageNet. It can be seen from From Table 4, our method significantly surpasses other
288 methods on Top-1 to Top-5 accuracy evaluations in bird, snake, primate and dog categories. Specially, on
289 Top-1 accuracy evaluations, our method even achieves *2-3 times* increasing compared with the state-of-
290 the-art. It demonstrates the shared visual features boost the description ability of the knowledge graph that
291 helps to find more unseen classes. And shared visual features also play a greater role than normal semantic
292 features where the model can double the performance with only a small amount of shared visual features
293 added. On Top-10 and Top-20 accuracy evaluations, DGP and our method have the similar performances.

Table 4. Top-k evaluation in Detailed Group

Test set	Model	hit@k(%)				
		1	2	5	10	20
bird	GCNZ (Wang et al., 2018b)	0.2	0.5	0.9	1.8	3.1
	SGCN (Kampffmeyer et al., 2019b)	2.4	6.2	**13.5**	**21.1**	**30.6**
	DGP (Kampffmeyer et al., 2019b)	2.3	6.0	13.2	20.8	20.2
	ours	**4.1**	**6.9**	**13.5**	20.3	29.6
snake	GCNZ (Wang et al., 2018b)	0.2	1.8	4.8	9.9	19.0
	SGCN (Kampffmeyer et al., 2019b)	4.9	10.2	22.8	**34.8**	**49.6**
	DGP (Kampffmeyer et al., 2019b)	4.2	9.4	22.8	34.8	49.3
	ours	**6.8**	**12.5**	**23.0**	33.9	47.9
primate	GCNZ (Wang et al., 2018b)	0.1	2.2	4.5	6.7	21.7
	SGCN (Kampffmeyer et al., 2019b)	9.7	21.2	44.4	66.8	79.7
	DGP (Kampffmeyer et al., 2019b)	9.6	22.7	49.0	**68.6**	**81.1**
	ours	**13.1**	**24.9**	**49.3**	64.5	78.9
dog	GCNZ (Wang et al., 2018b)	0.1	2.0	4.4	0.9	14.4
	SGCN (Kampffmeyer et al., 2019b)	6.2	19.4	37.7	46.7	**58.0**
	DGP (Kampffmeyer et al., 2019b)	6.2	18.6	37.1	46.2	57.1
	ours	**17.7**	**25.3**	**37.8**	**46.8**	56.7

Ablation Study

As defined in Equation 1 and 2, the proposed $SVKG$ mainly consists of X_s, X_v and X_{aug}. Thus, four combinations, "X_s" only, "$X_s + X_v$", "$X_s + X_{aug}$" and "All", for ablation study are established. Indeed, "X_s" denotes the traditional knowledge graph while "All" represents the proposed knowledge graph $SVKG_{aug}$. The experiments are conducted also on the Detailed group (Table 2) to evaluate Top-1 to Top-20 accuracy. The experimental result is presented in Table 5. It is clearly observed that the Top-1 and Top-2 accuracy has significant improvements with the addition of X_s, X_v and X_{aug}. Specially, in dog category, the Top-1 accuracy has a very steep growth from 6.2% (X_s) to 17.7% (All). The reason is that dog is a well-known category and widely used everywhere. This leads the upstream feature extractor, EfficientDet, to work well and extract more accurate part visual features for $SVKG$ that significantly improves the Top-1 accuracy. A similar phenomenon is also observed in bird and primate categories. The study thus also reveals that the better performance of the upstream feature extractor, the better accuracy the proposed $SVKG$ might achieve. In a short, the study demonstrates the effectiveness of each proposed module in real-world zero-shot learning tasks.

In the cases from Top-5 to Top-20, the best performance is "$X_s + X_{aug}$". The influence of graph augmentation X_{aug} is weakened after two or three hops, and distant nodes are easy to overfit with the presence of part visual nodes. Therefore, in the long-distance prediction, the overall performance of the model decreases.

Discussions

For further discussion, we conducted an unseen class search compared with DGP on Detailed group. Some representative results are presented in Figure 5. Obviously, our method obtains better accuracy on unseen class searching than DGP. Interestingly, it can be seen that our method predicts normally more specific class labels than DGP does, such as "plover - sea" (ours - DGP), "Australian blacksnake - lyre snake", "grivet - old word monkey" and "toy spaniel - toy". This means more fine-grained information is learned in our model that can be used to predict more specific unseen classes.

Meanwhile, t-SNE was used to visualize knowledge graph features for initial graph, GCNZ graph, DGP graph and $SVKG$ graph on dog category. The orange-colored nodes represent seen classes and the blue-colored nodes represent unseen classes. Figure 6(a) is the initial graph, where the features are represented by word embedding. It can be found that the graph feature distribution is disorderly

Table 5. The ablation study.

Test set	model	hit@k(%)				
		1	2	5	10	20
bird	X_s	2.3	6.0	13.2	20.8	20.2
	$X_s + X_v$	2.9	6.3	13.3	20.8	30.6
	$X_s + X_{aug}$	3.0	6.6	**13.8**	**21.3**	**31.0**
	All	**4.1**	**6.9**	13.5	20.2	29.6
snake	X_s	4.2	9.4	22.8	34.8	49.3
	$X_s + X_v$	4.8	10.6	22.4	33.3	46.1
	$X_s + X_{aug}$	6.4	11.9	**24.1**	**35.8**	**49.7**
	All	**6.8**	**12.5**	23.0	33.9	47.9
primate	X_s	9.6	22.7	49.0	68.6	81.1
	$X_s + X_v$	10.1	22.9	48.6	64.2	80.7
	$X_s + X_{aug}$	12.0	24.2	**51.5**	**67.4**	**81.9**
	All	**13.1**	**24.9**	49.3	64.5	78.9
dog	X_s	6.2	18.6	37.1	46.2	57.1
	$X_s + X_v$	5.3	13.6	34.9	45.2	55.6
	$X_s + X_{aug}$	14.5	24.6	**38.9**	**47.6**	**57.0**
	All	**17.7**	**25.3**	37.8	46.8	56.7
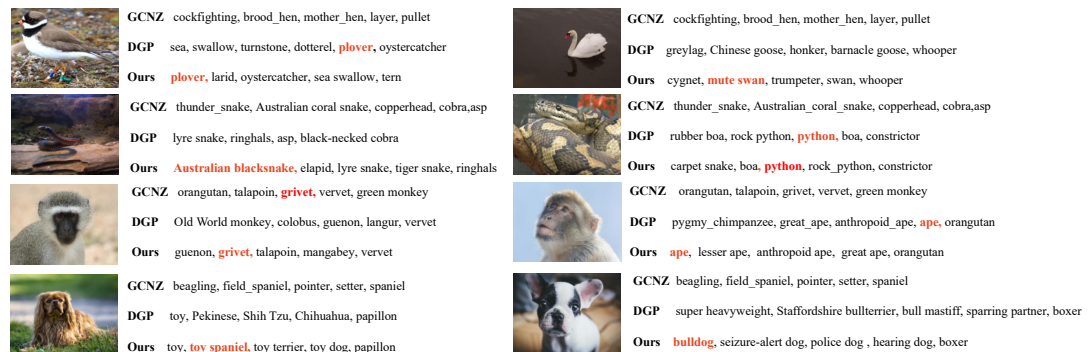
Figure 5. The Top-5 unseen classes searching compared with DGP. The correct label is colored and bold.

and the unseen classes are close to each other, which is not conducive for unseen class predicting. In GCNZ (Figure 6(b)) and DGP (Figure 6(c)), after cross-modal learning (semantic-to-visual modal), the graph features gradually become clear and show a hierarchical structure. Figure 6(d) shows the feature distribution of our graph $SVKG$. Intuitively, $SVKG$ is more structured and order than the graph features of GCNZ and DGP. It can be observed that blue-colored nodes distribute among orange-colored nodes reveals the hidden relationships between seen and unseen classes. This is because the proposed visual features X_v lead the graph representations of $SVKG$ close to real-world dog taxonomy. Unseen class nodes thus are distributed to the most related seen class nodes by the meaning of taxonomy. It obviously alleviates coincidence and proximity during unseen class predicting that improves the performance of zero-shot learning.

Moreover, we also compared feature distribution of $SVKG$ with GCNZ graph feature distribution and real image feature distribution, as shown in Figure 7. It can be observed that the feature distribution of $SVKG$ is significantly closer to the real image feature distribution than GCNZ. This indicates the feature of $SVKG$ is more close to the real visual feature that is easier to be matched with unseen classes. This also explains why the proposed method achieves the best accuracy on Top-1 and Top-2 unseen class predicting.

(a) Initial graph distribution   (b) GCNZ distribution   (c) DGP distribution   (d) Ours distribution
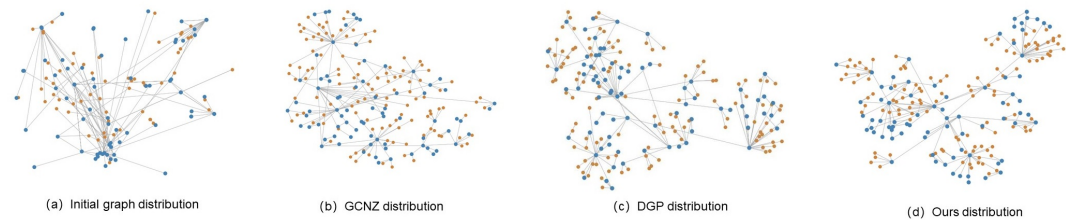
Figure 6. t-SNE visualizations for knowledge graph features of initial graph, GCNZ graph, DGP graph and ours *SVKG*.
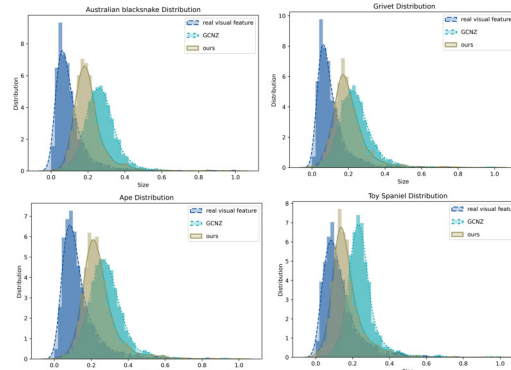
Figure 7. The feature distribution of real image feature, GCNZ graph feature, and *SVKG* feature.

CONCLUSION

In this paper, we propose a semantic-visual shared knowledge graph (*SVKG*) for zero-shot learning on the real-world dataset without needing pre-defined attributes. It combines semantic (from WordNet and Glove embedding) and visual features (extracted from raw images by EfficientDet) together in the same graph. The visual feature provides detailed information for describing fine-grained semantics that alleviates the "Domain-Shift" problem during the semantic-to-visual transformation of zero-shot learning. A novel multi-modal GCN model is also proposed to learn the graph representations of *SVKG*. After, the graph representations are further used for downstream zero-shot learning tasks in the experiments.

Experimental results on the real-world dataset demonstrate the effectiveness of our method and illustrate the multi-modal graph guide model generates more discriminative representation. And our method significantly surpasses other methods on Top-1 to Top-5 accuracy evaluations in the bird, snake, primate, and dog categories. Especially, on Top-1 accuracy evaluations, our method even achieves a 2-3 times increase compared with the state-of-the-art.

The important component of zero-shot learning tasks implemented in real-world environments is still how to reasonably use and construct the knowledge graph. In this paper, the *SVKG* is only storing fine-grained information in parts of visual features. In the future, we will add color, material, shape, and other relations and associated nodes to the *SVKG* in order to further increase the model's performance.

REFERENCES

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2013). Label-embedding for attribute-based classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 819–826.

Annadani, Y. and Biswas, S. (2018). Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7603–7612.

Cacheux, Y. L., Borgne, H. L., and Crucianu, M. (2019). Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10333–10342.

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., and Sun, X. (2020a). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *AAAI*, pages 3438–3445. AAAI Press.

Chen, S., Xie, G., Liu, Y., Peng, Q., Sun, B., Li, H., You, X., and Shao, L. (2021). HSVA: hierarchical semantic-visual adaptation for zero-shot learning. In *NeurIPS*, pages 16622–16634.

Chen, X., Lan, X., Sun, F., and Zheng, N. (2020b). A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference on Computer Vision*, pages 572–588. Springer.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129.

Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. (2015). Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2332–2345.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society.

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., and Xing, E. P. (2019a). Rethinking knowledge graph propagation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11487–11496.

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., and Xing, E. P. (2019b). Rethinking knowledge graph propagation for zero-shot learning. In *CVPR*, pages 11487–11496. Computer Vision Foundation / IEEE.

Kim, H., Lee, J., and Byun, H. (2021). Zero-shot learning with self-supervision by shuffling semantic embeddings. *Neurocomputing*, 437.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations,San Diego*.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations,Toulon France*.

Kodirov, E., Xiang, T., and Gong, S. (2017). Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3174–3183.

Lampert, C. H., Nickisch, H., and Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE.

Li, Y.-L., Xu, Y., Mao, X., and Lu, C. (2020). Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325.

Liu, L., Zhou, T., Long, G., Jiang, J., and Zhang, C. (2020). Attribute propagation network for graph zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4868–4875.

Liu, Y., Gao, X., Han, J., Liu, L., and Shao, L. (2022). Zero-shot learning via a specific rank-controlled semantic autoencoder. *Pattern Recognition*, 122:108237–.

Liu, Y., Guo, J., Cai, D., and He, X. (2019). Attribute attention for semantic disambiguation in zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6698–6707.

Liu, Y., Zhou, L., Bai, X., Huang, Y., Gu, L., Zhou, J., and Harada, T. (2021). Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3794–3803.

Min, S., Yao, H., Xie, H., Wang, C., Zha, Z.-J., and Zhang, Y. (2020). Domain-aware visual bias eliminating for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12664–12673.

Naeem, M. F., Xian, Y., Tombari, F., and Akata, Z. (2021). Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962.

Ni, J., Zhang, S., and Xie, H. (2019). Dual adversarial semantics-consistent network for generalized zero-shot learning. *Advances in Neural Information Processing Systems*, 32.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G., and Dean, J. (2014). Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations,Banff Canada*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL.

Rong, Y., Huang, W., Xu, T., and Huang, J. (2020). Dropedge: Towards deep graph convolutional networks on node classification. In *International Conference on Learning Representations,Addis Ababa Ethiopia*.

Song, W. and Zhang, L. (2022). Semantic-visual combination propagation network for zero-shot learning. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(4):2341–2345.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Tan, M., Pang, R., and Le, Q. V. (2020). Efficientdet: Scalable and efficient object detection. In *CVPR*, pages 10778–10787. Computer Vision Foundation / IEEE.

Wan, Z., Chen, D., Li, Y., Yan, X., Zhang, J., Yu, Y., and Liao, J. (2019). Transductive zero-shot learning with visual structure constraint. *Advances in Neural Information Processing Systems*, 32.

Wang, J. and Jiang, B. (2021a). Zero-shot learning via contrastive learning on dual knowledge graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 885–892.

Wang, J. and Jiang, B. (2021b). Zero-shot learning via contrastive learning on dual knowledge graphs. In *ICCVW*, pages 885–892. IEEE.

Wang, Q., Wu, W., Zhao, Y., and Zhuang, Y. (2021). Graph active learning for gcn-based zero-shot classification. *Neurocomputing*, 435.

Wang, X., Ye, Y., and Gupta, A. (2018a). Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866.

Wang, X., Ye, Y., and Gupta, A. (2018b). Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866. Computer Vision Foundation / IEEE Computer Society.

Xie, C., Xiang, H., Zeng, T., Yang, Y., Yu, B., and Liu, Q. (2021). Cross knowledge-based generative zero-shot learning approach with taxonomy regularization. *Neural Networks*, 139:168–178.

Xie, G.-S., Liu, L., Zhu, F., Zhao, F., Zhang, Z., Yao, Y., Qin, J., and Shao, L. (2020). Region graph embedding network for zero-shot learning. In *European conference on computer vision*, pages 562–580. Springer.

Xu, W., Xian, Y., Wang, J., Schiele, B., and Akata, Z. (2020). Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980.

Yu, H. and Lee, B. (2019). Zero-shot learning via simultaneous generating and learning. *Advances in Neural Information Processing Systems*, 32.

Zhao, T., Liu, Y., Neves, L., Woodford, O. J., Jiang, M., and Shah, N. (2021). Data augmentation for graph neural networks. In *AAAI*, pages 11015–11023. AAAI Press.

Zhu, P., Wang, H., and Saligrama, V. (2019). Generalized zero-shot recognition based on visually semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2995–3003.

Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., and Elgammal, A. (2018). A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76.