Exploring the Impact of Knowledge Graphs on Zero-Shot Visual Object State Classification

Filippos Gouidis[1,2][a], Konstantinos Papoutsakis[1][b], Theodore Patkos[3][c], Antonis Argyros[2,3][d] and Dimitris Plexousakis[2,3][e]

[1]*Department of Management, Science and Technology, Hellenic Mediterranean University, Agios Nikolaos, Greece*

[2]*Computer Science Department, University of Crete, Heraklion, Greece*

[3]*Institute of Computer Science, Foundation for Research and Technology Hellas, Heraklion, Greece*

Keywords: Visual Object State Classification, Zero-Shot Learning, Knowledge Graphs, Graph Neural Networks.

Abstract: In this work, we explore the potential of Knowledge Graphs (KGs) towards an effective Zero-Shot Learning (ZSL) approach for Object State Classification (OSC) in images. For this problem, the performance of traditional supervised learning methods is hindered mainly by data scarcity, as they attempt to encode the highly varying visual features of a multitude of combinations of object state and object type classes (e.g. open bottle, folded newspaper). The ZSL paradigm does indicate a promising alternative to enable the classification of object state classes by leveraging structured semantic descriptions acquired by external commonsense knowledge sources. We formulate an effective ZS-OSC scheme by employing a Transformer-based Graph Neural Network model and a pre-trained CNN classifier. We also investigate best practices for both the construction and integration of visually-grounded common-sense information based on KGs. An extensive experimental evaluation is reported using 4 related image datasets, 5 different knowledge repositories and 30 KGs that are constructed semi-automatically via querying known object state classes to retrieve contextual information at different node depths. The performance of vision-language models for ZS-OSC is also assessed. Overall, the obtained results suggest performance improvement for ZS-OSC models on all datasets, while both the size of a KG and the sources utilized for their construction are important for task performance.

1 INTRODUCTION

In recent years, the field of computer vision has witnessed remarkable advancements based on sophisticated Deep Neural Network models capable of performing various complex visual recognition tasks (Zhou et al., 2023). Traditional supervised learning methods exhibit state-of-the-art performance in various challenging problems based on labeled data for training, the collection and preparation of which is often expensive and time-consuming; a fact that hinders the application of the relevant methods in complex scenarios and open-world problems. Zero-shot Learning (ZSL) has emerged as a promising learning strategy to address this limitation (Xian et al., 2019). ZSL aims to enable learning of novel target classes not present in the training data by leveraging previously learned features as well as semantic descriptions or attributes, if available, that are associated with the classes (Lampert et al., 2013; Narayan et al., 2020). By exploiting features learned from the same or other datasets and knowledge transfer acquired by external data repositories from seen to unseen classes, ZSL provides a practical solution for recognizing the latter, thereby pushing the boundaries of visual recognition in challenging real-world scenarios (Monka et al., 2022; Pourpanah et al., 2023).

The more specific task of Zero-shot Object Recognition (ZSR) in images provides an intriguing extension of the ZSL paradigm, emphasizing the ability of machine learning models to generalize beyond the training samples. Such a strategy enables recognition of novel classes by integrating semantic attributes and their representation, i.e. in the form of feature embeddings, or textual descriptions associated with both known (seen) and novel (unseen) classes (Xian et al., 2019).

[a] https://orcid.org/0000-0002-9539-8749
[b] https://orcid.org/0000-0002-2467-8727
[c] https://orcid.org/0000-0001-6796-1015
[d] https://orcid.org/0000-0001-8230-3192
[e] https://orcid.org/0000-0002-0863-8266

Gouidis, F., Papoutsakis, K., Patkos, T., Argyros, A. and Plexousakis, D.
Exploring the Impact of Knowledge Graphs on Zero-Shot Visual Object State Classification.
DOI: 10.5220/0012434800003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 738-749
ISBN: 978-989-758-679-8; ISSN: 2184-4321

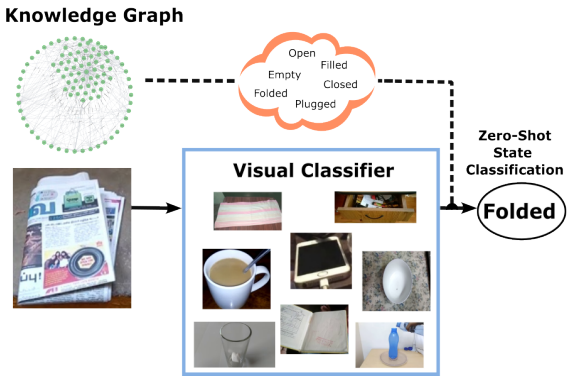


Figure 1: The proposed approach for Zero-shot Object State Classification combines structured representations of object states acquired from knowledge graphs with pre-trained visual information to infer previously unseen combinations of objects and object states.

In the pursuit of addressing the limitations of traditional supervised learning in Computer Vision, the integration of Knowledge Graphs (KGs) (Anh et al., 2021; Ilievski et al., 2021) also emerges as a promising line of research (Monka et al., 2022; Chen et al., 2023). General-purpose KGs contain domain, factual, and often commonsense knowledge by organizing semantic and possibly multi-modal features and relationships of entities, providing valuable encoding in symbolic form that can be integrated with neural models. In particular, annotation data from images or videos can be used to organize rich visually grounded knowledge into graphs using entities that are associated with various action types, human body parts, object classes and attributes, or other types of visual or non-visual information and their spatial or spatio-temporal relationships in case of video (Ghosh et al., 2020). By mining KGs for relevant semantic embeddings, ZSR models gain access to rich contextual knowledge, enabling a more efficient knowledge transfer between known and unknown/novel classes. Due to the immense potential of KGs in enriching visual recognition tasks, their role in this context is attracting increasing attention from researchers.

In the context of visual object recognition, object states can be viewed as a subset of perceptible object attributes. Attributes typically refer to static, inherent properties of objects, such as color, shape, or texture. In contrast, object states are associated with the dynamic aspects of changes in appearance, shape, and functionality (e.g., unfolded, closed, full etc.) which is related to a past action performed on the object (e.g., folding, closing, pouring etc.). Recognizing object states in images is generally more challenging compared to attributes due to the complexity involved in representing subtle visual information and contextual variations that object states entail. What is more, effective recognition of object states requires, according to the data-driven supervised learning paradigm, exhaustive training across a vast number of combinations of object classes and state classes, to capture their huge intra- and inter-class variability.

In this work, we aspire to investigate possible solutions for the task of visual Object State Classification (OSC) in images inspired by the paradigm of Zero-shot Classification using Knowledge Graphs (ZS-KG) (Nayak and Bach, 2022; Kampffmeyer et al., 2019). To achieve this goal (see Figure 1), we explore the construction of KGs and the integration of semantic information into Graph Convolutional Network models (GCNs), as powerful tools for learning visually-grounded knowledge in the context of ZSL. An effective CNN-based object classifier is also employed and adapted for ZS-OSC. The extensive experimental evaluation conducted suggests that learning structured semantic representations of the relationships among different objects and object state entities/concepts mined from KGs enables transfer learning to the CNN-based classifier with high accuracy. Thus, our main contributions are the following:

1. We formulate a ZSL approach for the task of OSC in images using KGs and GCN models. In contrast to existing methods (Gouidis et al., 2023), our work explores the more challenging, zero-shot variant of this task.
2. Multiple different KGs have been constructed to organize structured semantic information related to object states. We conduct a comparative study of their performance, as well as a comparison with Large Language Vision models toward the ZS-OSC task.
3. Our findings demonstrate improved performance toward the ZS-OSC task and the importance of using visually grounded KGs to enable the transfer of structured semantic knowledge related to object states into a deep neural classification model.

The project code/material is publicly available [1].

2 RELATED WORK

Object State/Attribute Recognition. The term "visual attributes" commonly refers to visual concepts that are perceivable by humans and AI-enabled agents (Duan et al., 2012). Currently, the prevalent approach for learning attributes is similar to

[1] https://github.com/papoutsakos/interlink

that of object categories, involving training convolutional neural networks with discriminative classifiers on annotated image datasets (Singh and Lee, 2016). Few works focus on state classification (Gouidis et al., 2022), while most of them rely on the same assumptions used for the attribute classification task. Recently, a multi-task, self-supervised learning method (Souček et al., 2022) was proposed to jointly learn to temporally localize object state changes and the corresponding state-modifying actions in videos. A prominent research direction to tackle this task refers to zero-shot learning that gained considerable attention in recent years due to its practical significance in real-world applications, mitigating the problem of collecting and learning training data for a very large number of object classes (Xian et al., 2018a). One prevalent zero-shot learning approach involves the use of semantic embeddings to represent objects and their attributes in a low-dimensional space (Wang et al., 2018).

Recently, the advent of powerful generative models also provided a promising research direction towards zero-shot object classification (Xian et al., 2018b; Changpinyo et al., 2016), by generating images of objects that resemble instances from seen/-known object classes. This enables the generation of new samples for previously unseen object classes. In the same line of work, the recent work by (Saini et al., 2023) focuses on the recognition of object states based on the concept of compositional generation of novel object-state images, also introducing the Chop & Learn dataset. In addition, recent studies have explored the potential of knowledge graphs in zero-shot learning (Kampffmeyer et al., 2019; Nayak and Bach, 2022).

Graph Neural Networks. Graph Neural Networks (GNNs) have become increasingly popular because of their capacity to learn node embeddings that capture the graph's structure (Kipf and Welling, 2016). These networks have demonstrated significant advancements in downstream tasks like node classification and graph classification (Hamilton et al., 2017; Wu et al., 2019; Vashishth et al., 2020). Previous studies have primarily viewed transformers as a means to learn meta-paths in heterogeneous graphs, rather than a technique for neighborhood aggregation. Additionally, GNNs have found applications in diverse areas, such as fine-grained entity typing (Xiong et al., 2019), text classification (Yao et al., 2019), reinforcement learning (Adhikari et al., 2020), and neural machine translation (Bastings et al., 2017). In our research, we employ a Transformer-based Graph Convolutional Network (GCN) model, which has recently been utilized in the context of zero-shot object classification (Nayak and Bach, 2022).

Common Sense Knowledge Graphs. Knowledge Graphs (KGs) can encode auxiliary semantic common-sense information through either a graph-based schema or a knowledge graph embedding represented in vector form (Bosselut et al., 2019). This important feature has recently attracted researchers to investigate numerous open-access Knowledge Graphs (KGs) that encompass universal information in conjunction with large vision datasets. Those KGs can serve as auxiliary knowledge in various vision-based problems.

Visualsem is a large, multi-modal KG for vision and language (Alberts et al., 2020) that incorporates multilingual information and visually grounded relations of entities, constructed using different publicly available knowledge sources (e.g., Wikipedia, ImageNet (Russakovsky et al., 2015), BabelNet v4.0 (Navigli and Ponzetto, 2012)). The VisionKG framework (Anh et al., 2021; Trung et al., 2021) [2] integrates labeled data across different, heterogeneous sources and computer vision datasets, such as the Visual Genome (Krishna et al., 2017), COCO, and KITTI. In (Giuliari et al., 2022) a heterogeneous Spatial Commonsense Graph is introduced for an effective integration between the commonsense knowledge and the spatial scene to efficiently tackle the task of graph-based object localization in partial scenes.

The CommonSense Knowledge Graph (CSKG) (Ilievski et al., 2021) is a large-scale, hyper-relational graph that combines seven popular sources of semantic information into a consolidated representation, such as: ConceptNet (Speer et al., 2017), Visual Genome, Wikidata (Vrandečić and Krötzsch, 2014) and WordNet (Miller, 1995), among others. It relies on the KGTK data model and file specification. Overall, KGs have been extensively employed successfully in various tasks including object classification (Zhang et al., 2019; Kampffmeyer et al., 2019; Xian et al., 2018a) and visual transfer learning (Alam et al., 2022; Bhagavatula et al., 2019).

Large Pre-Trained Models. Large Pre-trained Models (LPMs) constitute a special type of Large Language Models (LLMs) that exploit the idea of contrastive learning in order to achieve alignment between image and text. LPMs can be considered as an adaptation of the LLMs, which consists on training on massive amounts of text data, to computer vision tasks. More in detail, the typical approach behind LPMs is to train jointly an image encoder and a text encoder on millions of image-text pairs collected from internet. This allows the encoders to be able

[2] https://github.com/cqels/vision

to perform well on downstream tasks such as Image Captioning, Visual Question Answering and Zero-Shot Classification. Some typical examples of LPMs include CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021) and BLIP (Li et al., 2022).

Datasets. A set of publicly available image datasets that are linked with KGs also contain rich annotation data related to object states/attributes. Visual Genome (Krishna et al., 2017) is a large-scale dataset, particularly designed for tasks related to image classification and captioning, visual question answering and object recognition, among others, containing over $100K$ images and rich visually grounded annotation data for a wide variety of real-world scenarios. The Visual-Attributes-in-the-Wild (VAW) dataset (Pham et al., 2021)[3] is a large-scale image dataset providing explicitly positive and negative labels of visual object attributes related to appearance (color, texture), geometry (shape, size, posture), and other intrinsic object properties (state, action). Finally, the Object State Detection Dataset (OSDD) (Gouidis et al., 2022) provides more than $13K$ images and $19K$ annotations for 18 object categories and 9 state classes, namely open, close, empty, containing something liquid (CL), containing something solid (CS), plugged, unplugged, folded, and unfolded, based on the something-something V2 video dataset (Goyal et al., 2017).

3 METHODOLOGY

We formulate a ZSL approach for the task of OSC in images inspired by works that address the generalized ZS object or state classification problem(Kampffmeyer et al., 2019; Nayak and Bach, 2022; Gouidis et al., 2023). The main idea behind this line of work is that given a set of seen classes, the necessary information for the classification of the unseen classes can be found in a Knowledge Graph (KG), if processed appropriately by a Graph Convolutional Network (GCN). We aim to tackle the ZS task variation where the whole set of object state classes is considered previously unseen. An overview of the proposed approach is illustrated in Figure 2.

Let I_S denote a collection of images for which annotation data related to a set O_S of object-state classes is available. We assume a visual object classifier that is pre-trained according to a set of object classes O_C. Therefore, a visual feature vector $v_c \in \mathbb{R}$ of P dimensions is available for each $c \in O_C$. Moreover, a semantic representation is available for each class $s \in O_S$ and $c \in O_C$ as a word embedding $x \in \mathbb{R}^D$ of D dimensions, based on a KG, noted as train-KG, that is supported by the GloVe text model and word embeddings (Pennington et al., 2014).

Based on this information, we define a set of training data points acquired by the train-KG, noted $T_{KG} = (x_c, c)$, each containing a word embedding x_c for an object class $c \in O_C$, which is utilized for training a Graph Convolutional Neural Network model. Finally, we define a set of test data points as $T_{te} = \{x_s, s\}$ that are utilized to construct a task-specific KG, noted as OS-KG, which encodes structured semantic representations of all classes in O_C.

The goal of the proposed ZS-OSC approach is to adapt the pre-trained visual object classifier (OC) by leveraging the graph embeddings of the OS-KG model to replace the former's feature extraction layer. This process enables the visual classifier to infer the object state $s \in O_S$ in an image $I_i \in I_S$, regardless of the class $c \in O_C$ of the object that is present. We investigate different options related to the query node hop distance, the size, and the relation types for constructing the OS-KG and its role in achieving this goal, as described in the following. We employ a CNN-based classifier and assess its performance using as I_S four different datasets that provide annotation data for object states in images.

3.1 The Proposed ZS-OSC Pipeline

The pipeline of the method (Figure 2), comprises four stages:

1. Given a commonsense Knowledge Graph[4] and the corresponding GloVe features (Nayak and Bach, 2022; Pennington et al., 2014), a GCN model is trained to map its output word embeddings to the visual embeddings of a pre-trained CNN-based object classifier.
2. We construct a task-specific curated KG, noted OS-KG, using queries related to the object state classes O_S.
3. We use the GCN model on the OS-KG to obtain graph node embeddings for each class in O_S.
4. The set of graph embeddings is used to replace the feature extraction layer of the pre-trained OC, which enables it to infer any object state class in O_S given an input image, regardless of the object class represented.

This process enables generalizability and transfer learning, adapting the visual object classifier as a visual object state classifier. Similar to (Nayak and

[3]https://vawdataset.com/

[4]https://github.com/yinboc/DGP

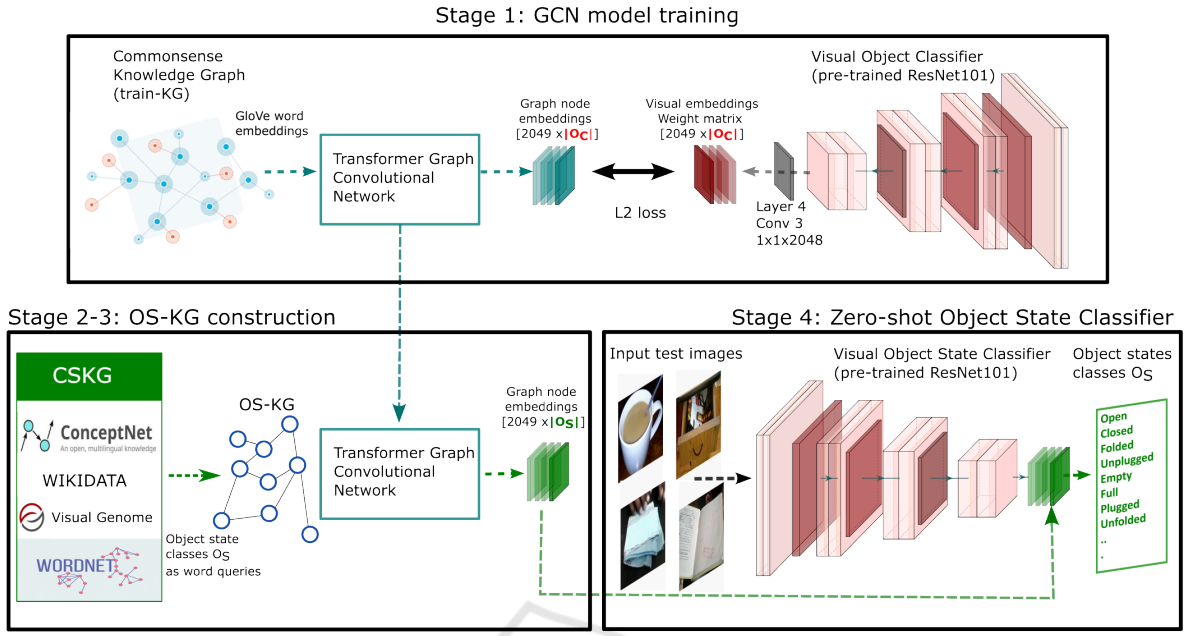


Figure 2: An overview of the proposed Zero-Shot Object State Classification approach.

Bach, 2022), we use the ConceptNet repository and the GloVe model (Pennington et al., 2014) to obtain word feature embedding vectors that are utilized for the GCN model training (Stage-1). We utilize the popular and effective ResNet101 model that is pre-trained as an OC classifier using $1K$ target object classes of the Imagenet (ILSVRC) dataset.

3.2 GCN Model Training

Graph Neural Networks are characterized by the capacity to encode the structure of a graph and the corresponding relationships between its nodes. This characteristic enables the learning of graph node embeddings by iterative aggregation of all k-hop neighbors of each graph node (Hamilton et al., 2017). The concept of Graph Convolutional Network model was originally proposed in (Kipf and Welling, 2016). A layer of a GCN implements two main functions (Xu et al., 2019), aggregation and combination.

$$\alpha_{\upsilon}^{(l)} = \text{AGGREGATE}^{(l)}\left(\left\{\mathbf{h}_{u}^{(l-1)} \forall u \in \mathcal{N}(\upsilon)\right\}\right) \quad (1)$$

In Equation 1, $\mathbf{h}_{u}^{(l-1)}$ regards the node feature vector for the neighborhood $\mathcal{N}$ of node υ, while $\alpha_{\upsilon}^{(l)}$ regards aggregated node feature of the set of neighbors.

Any aggregated node is used as input to the following function to generate a node feature $\mathbf{h}_{u}^{(l)}$ for the l-th layer of the network model starting from an initial GloVe word feature vector $\mathbf{h}_{u}^{(0)} = \mathbf{x}_{\upsilon}$:

$$\mathbf{h}_{\upsilon}^{(l)} = \text{COMBINE}^{(l)}\left(\mathbf{h}_{u}^{(l-1)}, \alpha_{\upsilon}^{(l)}\right). \quad (2)$$

We follow the 2-layer TrGCN model and the graph propagation module that was proposed in (Nayak and Bach, 2022). The ConceptNet (Speer et al., 2017) is employed as the commonsense KG for training the TrGCN model, as it best suits the formulation of a ZSL framework for OSC using KGs.

We set the last layer of the GCN model to match the dimensions of the CNN-based classifier's features extraction layer, that is a weight matrix $[P \times |O_C|]$. Each column comprises a set of weights that can be interpreted as a class-specific object classifier. Consequently, given a set T_{KG} of semantic features (e.g.GloVe) and graph topology information acquired by the OS-KG as input, the GCN model training is performed by minimizing the L2-distance classification loss between the weights of the semantic representations of the KG structure and the visual object classifier's weights. The train-KG combines semantic information for the classes of both sets of object classes O_C and object state classes O_S, as nodes, and their relationships as weighted edges, which is a key aspect of the proposed methodology. By using the features of the pre-trained CNN-based classifier for supervision, the GCN model is trained to generate graph embeddings using the train-KG, which implicitly encodes their relationships and embeds those into the visual feature space of the classifier.

Table 1: Type of data contained in the sources utilized for the construction of OS-KGs. CS: Common Sense. LX: Lexicographic. TH: Thesaurus. LD: Linked Data. AR: Affordance Related. IS: Image-Centric. SU: Scene Understanding LG: Logical.

KG	Relation Types
ConceptNet (CN)	CS, LX
WordNet (WN)	LX, TH
Wikidata (WK)	AR, LD
Visual Genome (VG)	IS, SU
CSKG	CS, LX, TH, LD, IS, SU, LG

3.3 KG Construction and Zero-Shot State Classification

For each object state class $c \in O_S$ we query the repository containing the topological and semantic information and retrieve a sub-graph that comprises the corresponding graph node, its k-hop neighborhood nodes, where $k = 1, 2, 3$. Each retrieved sub-graph is integrated into the corresponding OS-KG, while identical nodes are merged. The GCN that was trained in the previous stage is then utilized to obtain the graph embeddings for the OS-KG, which constitutes a $d \times |S|$ weight matrix and is used to replace the feature extraction layer of the CNN classifier.

For example, using the state class *open* as a query for $k = 1$ in Visual Genome, yields a large sub-graph with several nodes, e.g., *bottle*, *box*, *newspaper*, *book*, *jar* and *laptop*. It should be noted that the queries employed consider the existence of a relation and not its type. The query concerns whether two concepts are connected by any relation. In case those are connected, the corresponding nodes should also be connected in the sub-graph.

At inference time, a test image I_i is used as input to the adapted CNN classifier that is now able to estimate a visual feature vector for each object state class $c \in O_C$. The minimum L2 distance between the estimated visual features and the graph embedding f_s is calculated to finally classify the state of the object that is present in I_i, regardless of its object class.

The visual classifier demonstrates versatility, being capable of classifying the object state classes. This makes it suitable for zero-shot classification scenarios, extending its usability beyond traditional settings toward real-world applications and scenarios.

4 EXPERIMENTAL EVALUATION

We conducted a series of experiments to investigate the impact of the KG on the performance of our model. We construct several KGs, and each of them is experimentally assessed in our framework. The differences among the KGs refer to the source(s) to retrieve information and also to the hop node depth. Regarding the sources, we utilize 5 popular repositories and KG: ConceptNet (Speer et al., 2017), WordNet (Fellbaum, 2010), Wikidata (Vrandečić and Krötzsch, 2014),Visual Genome (Krishna et al., 2017) and CSKG (Ilievski et al., 2021). We also employ three depth levels for node search: hop k=1 to 3.

Regarding the knowledge sources utilized, information that is worthy of remark follows. ConceptNet offers a wide array of relational knowledge, capturing meaningful connections between various concepts extracted from a vast range of data sources. WordNet is a lexical database that contributes an extensive set of synsets, representing word meanings and their associations, thus bolstering the semantic depth of our KG. Wikidata, as a knowledge base of structured data, provides rich information about entities,

Table 2: The size and the relation types of each graph variant of OS-KG (rows) are reported. Each variant OS-KG has been constructed using a single or a combination of the available knowledge sources (CN: ConceptNet, VG:Visual Genome, WN:WordNet, WK:Wikidata, CSKG: Commonsense Knowledge Graph). H: Hop distance taken to construct the KG, Size of KGs. N: Number of Nodes. E: Number of Edges. RT: Number of Different Relation Types.

KG	H	N	E	RT
	1	821	1666	20
	2	27233	197950	39
CN	3	258603	6394846	47
	1	1018	2292	34
	2	14562	211974	6528
VG	3	25465	1851510	11811
	1	821	1,666	20
	2	27,233	197,950	39
CN,WK	3	25,8603	6,394,846	47
	1	1,018	2,292	34
	2	14,511	209,190	6,496
VG,WK	3	25,412	1,820,490	11,820
	1	1,031	2,314	35
	2	16,749	229,222	6,500
VG,WN	3	35,967	2,367,130	11,835
	1	1,834	3,958	53
	2	44,629	434,302	6,535
CN,VG,WK	3	300,426	9,184,186	12,046
	1	1,031	2,314	35
	2	16,749	229,222	6,500
VG,WK,WN	3	35,967	2,367,130	11,835
	1	1,845	3,980	53
	2	44,693	434,804	6,537
CN,VG,WN	3	300,867	9,200,084	12,048
	1	1,845	3,980	53
	2	44,693	434,804	6,537
CN,VG,WK,WN	3	300,867	9,200,084	12,048
	1	3,160	6,974	60
	2	103,391	993,782	6,560
CSKG	3	600,457	24,738,974	12,070

attributes, and their interconnections, thereby enhancing the KG with structured and linked data. Visual Genome and Common Sense Knowledge Graph add a multimodal dimension to our knowledge representation. Visual Genome, as a rich image-centric dataset, augments the KG with visual concepts and spatial relations extracted from images, bridging the gap between visual and textual knowledge. The Common Sense Knowledge Graph (CSKG) provides structured knowledge representation that captures general and domain-agnostic knowledge about the world incorporating all the aforementioned knowledge sources, among others, into a large-scale knowledge repository. Overall, we conducted experiments using 30 KG variants generated based on different combinations of sources and node search depths/hops. The details regarding the KGs are shown in Table 1 and Table 2.

4.1 Implementation & Evaluation Issues

Implementation Details. We employ the ImageNet-based KG (Kampffmeyer et al., 2019)[5] as the train-KG model. The GCN model was trained from scratch following the methodology outlined in (Giuliari et al., 2022) and in (Nayak and Bach, 2022). The training process involved 1000 epochs using 950 randomly selected classes from the ImageNet (ILSVRC 2012) dataset (Russakovsky et al., 2015), while the remaining 50 classes were reserved for validation. The GCN model with the lowest validation loss was selected to generate embeddings for both object and object state classes using the test KG.

Datasets. Except for the OSDD dataset (Gouidis et al., 2022), which is specifically designed for state detection, there is no other dataset that focuses exclusively on object states in images, at this moment. However, some existing object detection and classification datasets include object states as a subset of their object classes. These include the Visual Attributes in the Wild (VAW) dataset (Pham et al., 2021) that includes object state classes as a subset of attribute annotations. Likewise, MIT-States (Isola et al., 2015) and CGQA-States (Mancini et al., 2022) are two widely used datasets used in the context of attribute classification. To leverage VAW, MIT-States, and CGQA-States for our experimental evaluation, we extracted subsets that specifically pertain to object states. Additionally, for the OSDD and VAW datasets, we extracted the bounding boxes from the original images to create suitable images for the OSC task. A simple analysis that reveals the complexity of each dataset is to consider (i) the number of the target state classes and (ii) the average number of states per object class (a higher ratio typically corresponds to greater complexity), as reported in Table 3.

Metrics. Our evaluation protocol adheres to the standard zero-shot evaluation method as described in (Purushwalkam et al., 2019). In contrast to the standard setting where the accuracy over all classes is reported, in this case after the accuracy for each class is computed, an overall mean average across the previous results is reported. This approach treats each class equally since it does not take into account the corresponding number of samples of each class.

Competing Methods. To our knowledge, currently there is no object-agnostic state model that can be used off-the-shelf in the context of zero-shot setting. Therefore we opt to use three SoA LPMs: CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021) and BLIP (Li et al., 2022), which support this functionality. Overall, we experiment with two version of CLIP and one version of ALIGN and BLIP respectively. It should be noticed that all of these models violate indirectly the basic assumptions of zero-shot setting since the pairs of text and images that have been used for the training contain the target classes for our task.

4.2 Results

Table 4 presents a comprehensive evaluation of various Knowledge Graphs (KGs) and language-vision models across four different image datasets that are either designed for the task of object state classification or include augmented annotation data related to object states. Regarding the models of 1 hop, **C-WK-WN** and **CN-WN** perform best in OSDD, **VG** variant excels in the CGQA-states and VAW dataset, while the **CN-VG-WN** model achieves the highest performance in the MIT-States. Concerning the model constructed using 2 hops, **C-VG-WN** achieves the highest performance in the OSDD and the MIT-States, while **VG** exhibits the top performance in CGQA-states. Finally, in the case of the 3 hops, **C-WK-WN** and **CN-WN** are the best variants in OSDD, **VG-WN** and **VG-WK-WN** are the best variants in CGQA-States, **CN-VG** and **CN-VG-WK** are the best variants in MIT-States and **VG** is the best model in the VAW respectively. Overall, the **VG** exhibits the best performance with 4 top performances across the 12 comparisons (4 different datasets $\times$ 3 different hops).

A closer examination of these outcomes reveals that models constructed using the same KG for depth of hop $k = 1$ hop or $k = 2$ outperform those constructed using $k = 3$ in most cases. This observation suggests that further augmenting the KG beyond

[5]Publicly available at https://github.com/yinboc/DGP.

Table 3: We report details on the four image datasets utilized in this work. Train/Val/Test: Number of Training/Validation/Testing Images. States: Number of State classes, Objects: Number of Object classes. VOSC/TOSC: Valid/Total Object-State combinations. S\O: Average number of states than an Object can be situated in.

Dataset	Train	Val	Test	States	Objects	VOSC	TOSC	S\O
OSDD (Gouidis et al., 2022)	6,977	1,124	5,275	9	14	35	126	2.36
CGQA-states (Mancini et al., 2022)	244	46	806	5	17	41	75	1.71
MIT-states (Isola et al., 2015)	170	34	274	5	14	20	70	1.57
VAW (Pham et al., 2021)	2,752	516	1,584	9	23	51	207	2.61

Table 4: Experimental results of the proposed approach for the zero-shot object state classification task. The reported scores summarize the average accuracy scores (columns) in the form of triplets for the hop-1 / hop-2 / hop-3 node depth/distance options for each dataset (columns). Each row represents the performance obtained by using a different OS-KG model that is manually constructed using the combination of the reported knowledge sources. Additionally, the performance of vision-language models is reported as well as that of a supervised visual state classification model, as reference. The latter relies on the ResNet-101 network model that is trained in a fully supervised setting on each dataset separately. VG: Visual Genome-based model. CN: ConceptNet-based model. WN: WordNet-based model. WK: Wikidata-based model. CSKG: Common-sense Knowledge Graph-based model (incorporates all other knowledge sources). The performance for four datasets, OSDD (Gouidis et al., 2022), CGQA-States (Naeem et al., 2021), MIT-States (Isola et al., 2015), VAW (Pham et al., 2021), is reported. Bold and underlined scores indicate the best performance across category and among all methods, respectively.

Knowledge Graph/Model	OSDD	CGQA-States	MIT-States	VAW
CN-VG-WK-WN	28.3 / **28.4** / 26.4	42.7 / 42.0 / 41.0	39.3 / 39.4 / 33.0	22.4 / 22.6 / 18.8
CN-VG-WK	25.7 / 25.6 / 26.4	42.4 / 43.6 / 40.0	34.7 / 34.8 / **36.0**	21.1 / 21.0 / 18.3
CN-VG-WN	28.3 / **28.4** / 26.4	42.7 / 42.0 / 41.0	**39.3 / 39.4 /** 33.0	22.4 / 22.5 / 18.8
CN-VG	25.7 / 25.6 / 24.4	42.4 / 43.6 / 40.0	34.7 / 34.8 / **36.0**	21.1 / 21.0 / 18.3
CN	26.3 / 26.3 / 26.9	40.7 / 40.7 / 42.4	35.0 / 35.0 / 31.7	21.6 / 21.7 / 20.7
VG-WK-WN	**29.1** / 27.2 / **27.6**	43.3 / 43.8 / **42.8**	36.2 / 38.2 / 35.7	23.9 / 23.5 / 22.7
VG-WK	26.9 / 27.3 / 25.2	46.7 / **47.4** / 37.2	38.6 / 39.3 / 34.2	**25.4** / 23.9 / 24.9
VG-WN	**29.1** / 27.2 / **27.6**	43.3 / 43.8 / **42.8**	36.2 / 38.2 / 35.7	24.0 / 23.4 / 22.7
VG	26.9 / 27.3 / 25.2	**46.7 / 47.4** / 37.2	38.6 / 39.2 / 34.2	**25.4** / 23.9 / **24.9**
CSKG	28.3 / 28.1 / 26.0	40.0 / 40.0 / 44.0	38.1 / 38.1 / 35.5	21.5 / **24.4** / 21.9
CLIP-RN101	22.5	**46.9**	39.3	28.0
CLIP-VT16	28.8	44.9	**46.4**	**30.1**
ALIGN	**29.5**	40.0	44.2	28.4
BLIP	13.3	26.0	27.2	16.1
Supervised State Classifier	67.5	60.5	85.3	51.9

a certain size yields no additional benefits and may introduce noise that deteriorates model performance.

Regarding the KG construction sources, most of the top-performing models either include Visual Genome (VG) or are based solely on it. These models consistently rank among the top positions across all datasets, demonstrating the robustness and potential of this visual-centric dataset. Conversely, models built using the greatest number of sources, such as **CSKG** and **VG-CN-WN-WK** exhibit rather mediocre performance, possibly due to overlapping information and susceptibility to noisy data present in these KGs. Likewise, the model based solely on ConceptNet (CN) ranks as one of the worst models across most of the comparisons. Finally, another interesting finding concerns the fact the model consisting of Visual Genome and ConceptNet, performs in most cases worst than the corresponding models consisting solely of either Visual Genome or ConceptNet.

Concerning the results obtained based on the LPMs, the CLIP-VT16 exhibits the best performance in MIT-States and VAW, CLIP-RN101 in CGQA-States and ALIGN in OSDD respectively. Except for the CGQA-states, the obtained results outperform the results obtained by the top OS-KG models. However, two important factors should be taken into consideration. First, the considerably larger training set used to train the visual backbones of LMPs that is orders of magnitude greater in comparison to the amount of the OS-KG models[6]. Moreover, LPMs have en-

[6]CLIP was trained approximately on 4×10^8 images and ALIGN on about 8×10^8. The backbone of the OS-KGs models used approximately 1×10^6 images.

Figure 3: Qualitative results of the proposed ZS-OSC approach using images from the OSDD dataset. The VG-WN knowledge sources is used to generate the OS-KG model in this case. For each sample image, the predicted object state class and the ground truth class labels are noted. Both correct (top row) and incorrect (bottom row) predictions are illustrated.

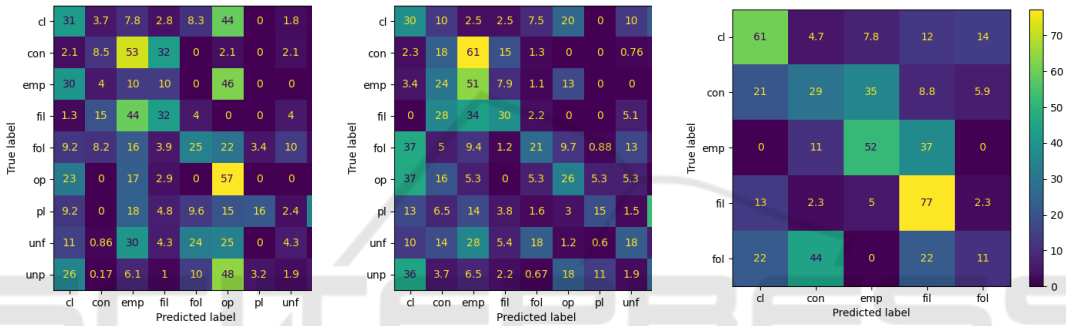


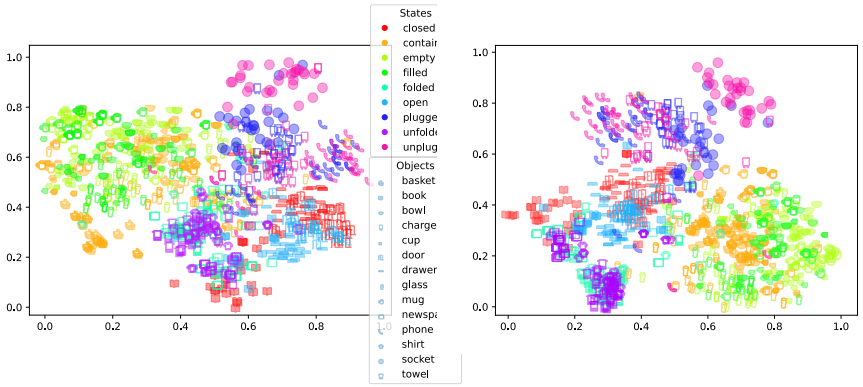
Figure 4: Confusion matrices of the model based on the VG-WN KG for OSDD, VAW and CGQA-states dataset. The numbers reported are % percentages of correct predictions. (cl: Closed, con: Containing, emp: Empty, fil: Filled, fol: Folded, op: Open, pl: Plugged, unf: Unfolded, unp: Unplugged).

countered during the training text-images pairs corresponding to the target classes. Finally, if we focus on CLIP-RN101 which is the only LPM that uses the same visual backbone as the OS-KG modes, we observe that it is outperformed by all OS-KG models in the OSDD dataset and by two OS-KG models in the CGQA-States dataset, respectively.

Based on these observations, it becomes evident that a node inclusion policy, in addition to the hop depth criterion, could enhance model performance. Furthermore, a more sophisticated approach is necessary to effectively combine different sources, considering information overlap and complementarity, thereby mitigating noise and further improving model accuracy and generalizability. These insights pave the way for future research, aiming to optimize KG-based models for zero-shot object state classification tasks.

A set of qualitative results is also illustrated in Figure 3, using the proposed ZS-OSC approach and the VG-WN knowledge sources to generate the OS-KG model. Both correct (top row) and incorrect (bottom row) predictions are shown, revealing some of the challenges an efficient solution to the OSC task has to deal with. Estimating the object state class regardless of the actual class/type of the object that is present considerably hinders the performance of appearance-based approaches that need to encode the large appearance variability of objects from different classes that possibly share the same state class, i.e. open drawer vs open bottle vs open door, and the subtle perceptible changes in the appearance of similar objects that constitute its current state, i.e. closed vs open bottle. The appearance of objects or background image content deteriorates the model's performance, i.e. we speculate that the human finger that appears to touch the smartphone's edge in the image of the last row and column in Figure 3 resembling a cable, causes the model to misclassify the object state as plugged, while a transparent bowl placed upside-down is mistaken for a filled container.

Similar conclusions can be drawn by the examination of the confusion matrices of the different mod-



(a) (b)

Figure 5: t-SNE visualization of visual features extracted from images of the OSDD dataset. The visual features are generated in (a) using the visual object classifier that is pre-trained on the ImageNet dataset and in (b) using the Supervised State Classifier that has been fine-tuned on the dataset. Samples are illustrated in different colors to represent the nine target object state classes of the OSDD dataset. Different marks are used to represent the fourteen object classes.

els, where it can be seen that the wrong predictions correspond mainly to states related to the gt state. In the case of the OSDD and VAW the related states are grouped in 3 pairs (closed-open, folded-unfolded, plugged-unplugged) and one triplet (empty-containing-filled), while in the case of the OSDD and VAW there is only the triplet of related states(empty-containing-filled). The confusion matrices that are produced by the model based on the VG-WN KG for three datasets are shown in Figure 4.

Another aspect of the approach that merits highlighting is the contribution of the object classifier, the weights of which are used by all but the last layer of the RN101 classifier that is used for the ZS-OSC. Specifically, the object classifier has been trained on the 1000 classes of the ImageNet dataset and, therefore, the weights of these layers can be considered as encapsulating the recognition of these object classes.

Those remarks can be further supported based on observations of the t-SNE visualization (Van der Maaten and Hinton, 2008) of the RN101 classifier features illustrated in Figure 5. Features extracted by two variants of the classifier were used (a) using the visual object classifier that is pre-trained on the ImageNet dataset and in (b) using the Supervised State Classifier that has been fine-tuned on the OSDD dataset. In Figure 5a, the t-sne output reveals discriminative clustering that is indicative of the groups of target classes that have been overlaid using distinct marks, as samples of the same object but different state classes tend to lie closer in the feature space than samples of the same state but different object class. In Figure 5b, the fine tuning appears to improve substantially the clustering mitigating this issue suggesting the important role of the object state classes.

5 CONCLUSIONS

In this study, we formulate a novel approach for the zero-shot object state classification (ZS-OSC) task using Knowledge Graphs (KGs) and extensively evaluate the effectiveness of various types of KGs. The comparative evaluation is conducted on four benchmark datasets (Gouidis et al., 2022; Krishna et al., 2017; Mancini et al., 2022; Pham et al., 2021). The results reveal an optimal threshold to be sought regarding the number of KG nodes. Beyond this threshold, including additional nodes leads to a decline in model performance, highlighting the importance of carefully selecting the KG size. Moreover, the type of knowledge encoded in the KG has a crucial role, as visually grounded semantic information appears more suitable to efficiently represent features and complex relations of semantic entities.

We argue that the zero-shot learning paradigm has great potential in improving the state-of-the-art performance for the OSC task by exploring intriguing future steps to extend our presented work, such as (a) fine-tuning of the GCN model using a visual classifier for attribute classes or object-attributes pairs, (b) integration of more powerful visual classifiers based on transformer models and (b) more elaborate techniques to construct visually grounded commonsense KGs and to fuse rich semantic information in deep neural models following the recent advancements of the vision-language models.

ACKNOWLEDGEMENTS

The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the 3rd Call for H.F.R.I. Research Projects to support Post-Doctoral Researchers (Project Number 7678 InterLinK: Visual Recognition and Anticipation of Human-Object Interactions using Deep Learning, Knowledge Graphs and Reasoning).

REFERENCES

Adhikari, A., Yuan, X., Côté, M.-A., Zelinka, M., Rondeau, M.-A., Laroche, R., Poupart, P., Tang, J., Trischler, A., and Hamilton, W. (2020). Learning dynamic belief graphs to generalize on text-based games. *N*IPS, 33:3045–3057.

Alam, M., Buscaldi, D., Cochez, M., Osborne, F., Reforgiato Recupero, D., Sack, H., Monka, S., Halilaj, L., Rettinger, A., Alam, M., Buscaldi, D., Cochez, M., Osborne, F., Refogiato Recupero, D., and Sack, H. (2022). A survey on visual transfer learning using knowledge graphs. *S*emant. Web, 13(3):477510.

Alberts, H., Huang, T., Deshpande, Y., Liu, Y., Cho, K., Vania, C., and Calixto, I. (2020). Visualsem: a high-quality knowledge graph for vision and language. *a*rXiv preprint arXiv:2008.09150.

Anh, L.-T., Manh, N.-D., Jicheng, Y., Trung, K.-T., Manfred, H., and Danh, L.-P. (2021). Visionkg: Towards a unified vision knowledge graph. In *P*roceedings of the ISWC 2021 Posters & Demonstrations Track, Workshop Proceedings.

Bastings, J., Titov, I., Aziz, W., Marcheggiani, D., and Sima'an, K. (2017). Graph convolutional encoders for syntax-aware neural machine translation. *a*rXiv preprint arXiv:1704.04675.

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S. W.-t., and Choi, Y. (2019). Abductive commonsense reasoning. *a*rXiv preprint arXiv:1908.05739.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (2019). Comet: Commonsense transformers for automatic knowledge graph construction. *a*rXiv preprint arXiv:1906.05317.

Changpinyo, S., Chao, W.-L., Gong, B., and Sha, F. (2016). Synthesized classifiers for zero-shot learning. In *P*roceedings of the IEEE CVPR, pages 5327–5336.

Chen, J., Geng, Y., Chen, Z., Pan, J. Z., He, Y., Zhang, W., Horrocks, I., and Chen, H. (2023). Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. *P*roceedings of the IEEE.

Duan, K., Parikh, D., Crandall, D., and Grauman, K. (2012). Discovering localized attributes for fine-grained recognition. In *2*012 IEEE CVPR, pages 3474–3481. IEEE.

Fellbaum, C. (2010). Wordnet. In *T*heory and applications of ontology: computer applications, pages 231–243. Springer.

Ghosh, P., Saini, N., Davis, L. S., and Shrivastava, A. (2020). All about knowledge graphs for actions. *a*rXiv preprint arXiv:2008.12432.

Giuliari, F., Skenderi, G., Cristani, M., Wang, Y., and Del Bue, A. (2022). Spatial commonsense graph for object localisation in partial scenes. In *P*roceedings of the IEEE/CVF CVPR, pages 19518–19527.

Gouidis, F., Patkos, T., Argyros, A., and Plexousakis, D. (2022). Detecting object states vs detecting objects: A new dataset and a quantitative experimental study. In *P*roceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), volume 5, pages 590–600.

Gouidis, F., Patkos, T., Argyros, A., and Plexousakis, D. (2023). Leveraging knowledge graphs for zero-shot object-agnostic state classification. *a*rXiv preprint arXiv:2307.12179.

Goyal, R., Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., and Memisevic, R. (2017). The something something video database for learning and evaluating visual common sense. In *2*017 IEEE ICCV, pages 5843–5851, Los Alamitos, CA, USA. IEEE Computer Society.

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. *N*IPS, 30.

Ilievski, F., Szekely, P., and Zhang, B. (2021). Cskg: The commonsense knowledge graph. *E*xtended Semantic Web Conference (ESWC).

Isola, P., Lim, J. J., and Adelson, E. H. (2015). Discovering states and transformations in image collections. *P*roceedings of the IEEE Computer Society CVPR, 07-12-June:1383–1391.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *I*CML, pages 4904–4916. PMLR.

Kampffmeyer, M., Chen, Y., Liang, X., Wang, H., Zhang, Y., and Xing, E. P. (2019). Rethinking knowledge graph propagation for zero-shot learning. *P*roceedings of the IEEE Computer Society CVPR, 2019-June:11479–11488.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *a*rXiv preprint arXiv:1609.02907.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *I*JCV, 123:32–73.

Lampert, C. H., Nickisch, H., and Harmeling, S. (2013). Attribute-based classification for zero-shot visual object categorization. *I*EEE Trans. on PAMI, 36(3):453–465.

Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *I*CML, pages 12888–12900. PMLR.

Mancini, M., Naeem, M. F., Xian, Y., and Akata, Z. (2022). Learning Graph Embeddings for Open World Compositional Zero-Shot Learning. *I*EEE Trans. on PAMI, 8828(c):1–15.

Miller, G. A. (1995). Wordnet: a lexical database for english. *C*ommunications of the ACM, 38(11):39–41.

Monka, S., Halilaj, L., and Rettinger, A. (2022). A survey on visual transfer learning using knowledge graphs. *S*emantic Web, 13(3):477–510.

Naeem, M. F., Xian, Y., Tombari, F., and Akata, Z. (2021). Learning graph embeddings for compositional zero-shot learning. In *P*roceedings of the IEEE/CVF CVPR, pages 953–962.

Narayan, S., Gupta, A., Khan, F. S., Snoek, C. G., and Shao, L. (2020). Latent embedding feedback and discriminative features for zero-shot classification. In *C*omputer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, pages 479–495. Springer.

Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *A*rtificial Intelligence, 193:217–250.

Nayak, N. V. and Bach, S. H. (2022). Zero-shot learning with common sense knowledge graphs. *T*rans. on Machine Learning Research (TMLR).

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *P*roceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., and Shrivastava, A. (2021). Learning to predict visual attributes in the wild. In *P*roceedings of the IEEE/CVF CVPR, pages 13018–13028.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., Wang, X.-Z., and Wu, Q. M. J. (2023). A review of generalized zero-shot learning methods. *I*EEE Trans. on PAMI, 45(4):4051–4070.

Purushwalkam, S., Nickel, M., Gupta, A., and Ranzato, M. (2019). Task-driven modular networks for zero-shot compositional learning. In *P*roceedings of the IEEE/CVF International Conference on Computer Vision, pages 3593–3602.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *I*CML, pages 8748–8763. PMLR.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *I*JCV, 115:211–252.

Saini, N., Wang, H., Swaminathan, A., Jayasundara, V., He, B., Gupta, K., and Shrivastava, A. (2023). Chop & learn: Recognizing and generating object-state compositions. In *P*roceedings of the IEEE/CVF ICCV, pages 20247–20258.

Singh, K. K. and Lee, Y. J. (2016). End-to-end localization and ranking for relative attributes. In *C*omputer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14, pages 753–769. Springer.

Souček, T., Alayrac, J.-B., Miech, A., Laptev, I., and Sivic, J. (2022). Multi-task learning of object state changes from uncurated videos.

Speer, R., Chin, J., and Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge. In *P*roceedings of the AAAI conference on artificial intelligence, volume 31.

Trung, K.-T., Anh, L.-T., Manh, N.-D., Jicheng, Y., and Danh, L.-P. (2021). Fantastic data and how to query them. In *P*roceedings of the NeurIPS 2021 Workshop on Data-Centric AI, Workshop Proceedings.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *J*ournal of machine learning research, 9(11).

Vashishth, S., Sanyal, S., Nitin, V., and Talukdar, P. (2020). Composition-based multi-relational graph convolutional networks. In *I*nternational Conference on Learning Representations.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *C*ommunications of the ACM, 57(10):78–85.

Wang, X., Ye, Y., and Gupta, A. (2018). Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs. *P*roceedings of the IEEE CVPR, pages 6857–6866.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. (2019). Simplifying graph convolutional networks. In *I*CML, pages 6861–6871. PMLR.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2018a). Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly. *I*EEE Trans. on PAMI, 41(9):2251–2265.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. (2019). Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly. *I*EEE Trans. on Pattern Analysis & Machine Intelligence, 41(09):2251–2265.

Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. (2018b). Feature generating networks for zero-shot learning. pages 5542–5551.

Xiong, W., Wu, J., Lei, D., Yu, M., Chang, S., Guo, X., and Wang, W. Y. (2019). Imposing label-relational inductive bias for extremely fine-grained entity typing. *ar*Xiv preprint arXiv:1903.02591.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2019). How powerful are graph neural networks? In *I*nternational Conference on Learning Representations.

Yao, L., Mao, C., and Luo, Y. (2019). Graph convolutional networks for text classification. In *P*roceedings of the AAAI conference on artificial intelligence, volume 33, pages 7370–7377.

Zhang, C., Lyu, X., and Tang, Z. (2019). Tgg: Transferable graph generation for zero-shot and few-shot learning. In *P*roceedings of the 27th ACM International Conference on Multimedia, pages 1641–1649.

Zhou, L., Meng, X., Liu, Z., Wu, M., Gao, Z., and Wang, P. (2023). Human pose-based estimation, tracking and action recognition with deep learning: A survey.