Semantic Web 13 (2022) 477–510
DOI 10.3233/SW-212959
IOS Press

A survey on visual transfer learning using knowledge graphs

Sebastian Monka [a,b,*], Lavdim Halilaj [a] and Achim Rettinger [b]
[a] *Corporate Research, Robert Bosch GmbH, Renningen, Germany*
E-mails: sebastian.monka@de.bosch.com, lavdim.halilaj@de.bosch.com
[b] *Computer Sciences, Trier University, Trier, Germany*
E-mail: rettinger@uni-trier.de

Abstract. The information perceived via visual observations of real-world phenomena is unstructured and complex. *Computer vision* (CV) is the field of research that attempts to make use of that information. Recent approaches of CV utilize *deep learning* (DL) methods as they perform quite well if training and testing domains follow the same underlying data distribution. However, it has been shown that minor variations in the images that occur when these methods are used in the real world can lead to unpredictable and catastrophic errors. Transfer learning is the area of machine learning that tries to prevent these errors. Especially, approaches that augment image data using auxiliary knowledge encoded in language embeddings or *knowledge graphs* (KGs) have achieved promising results in recent years. This survey focuses on visual transfer learning approaches using KGs, as we believe that KGs are well suited to store and represent any kind of auxiliary knowledge. KGs can represent auxiliary knowledge either in an underlying graph-structured schema or in a vector-based *knowledge graph embedding*. Intending to enable the reader to solve visual transfer learning problems with the help of specific KG-DL configurations we start with a description of relevant modeling structures of a KG of various expressions, such as directed labeled graphs, hypergraphs, and hyper-relational graphs. We explain the notion of feature extractor, while specifically referring to visual and semantic features. We provide a broad overview of *knowledge graph embedding methods* and describe several joint training objectives suitable to combine them with high dimensional visual embeddings. The main section introduces four different categories on how a KG can be combined with a DL pipeline: 1) Knowledge Graph as a Reviewer; 2) Knowledge Graph as a Trainee; 3) Knowledge Graph as a Trainer; and 4) Knowledge Graph as a Peer. To help researchers find meaningful evaluation benchmarks, we provide an overview of generic KGs and a set of image processing datasets and benchmarks that include various types of auxiliary knowledge. Last, we summarize related surveys and give an outlook about challenges and open issues for future research.

Keywords: Knowledge graph, visual transfer learning, knowledge-based machine learning

1. Introduction

Deep learning (DL) as a *machine learning* (ML) technique is broadly used to successfully solve *computer vision* (CV) tasks. Their main strength is their ability to find complex underlying features in a given set of images.

[*] Corresponding author. E-mail: sebastian.monka@de.bosch.com.

A common method for training a *deep neural network* (DNN) is to minimize the *cross-entropy* (CE) loss, which is equivalent to maximizing the negative log-likelihood between the empirical distribution of the training set and the probability distribution defined by the model. This relies on the *independent and identically distributed* (i.i.d.) assumptions as underlying rules of basic ML, which state that the examples in each dataset are independent of each other, that train and test set are identically distributed and drawn from the same probability distribution [47]. However, if the train and test domains follow different image distributions the i.i.d. assumptions are violated, and DL leads to unpredictable and poor results [131]. This has been demonstrated by using adversarially constructed examples [48] or variations in the test images such as noise, blur, and JPEG compression [55]. Moreover, authors in [26] even claim that any standard DNN suffers from such an unpredictable distribution shift when it is deployed in the real world.

Transfer learning is the area of machine learning that groups approaches dealing with such an unpredictable distribution shift [26]. Most of the transfer learning approaches try to solve the problem by inducing a bias into the DNN to overcome data issues. Especially, approaches that extend image data using auxiliary knowledge encoded in language embeddings or *knowledge graphs* (KGs) have achieved promising results in recent years. Due to Larochelle et al. [78] auxiliary knowledge is not only important to solve transfer learning problems, but also an opportunity to influence the way an ML model learns from unstructured data.

In this survey, we focus on visual transfer learning approaches using KGs, as we believe that KGs are well suited to store and represent any kind of auxiliary knowledge. The auxiliary knowledge encoded in an underlying graph-structured schema can then be converted to a vector-based *knowledge graph embedding* (h_s). The ability to transform the graph-based knowledge into the vector space enables the application of linear operations thus its use in combination with DNNs. A commonly used method for introducing auxiliary knowledge is to use a joint training objective that combines the semantic embedding h_s with the visual embedding h_v. In the scope of the survey we introduce three distinct types of joint embeddings: a) A semantic-visual embedding $h_{s,v}$, where semantic data is embedded using h_v as an objective; b) A visual-semantic embedding $h_{v,s}$, where visual data is embedded using h_s as an objective; and c) A hybrid embedding h_h, where both semantic and visual data are embedded using a combination of h_s and h_v as an objective.

Our main contributions in this survey are listed in the following:

- A categorization of visual transfer learning approaches using KGs according to four distinct ways a KG can be combined with a DL pipeline.
- A description of generic KGs and relevant datasets and benchmarks for visual transfer learning using KGs for CV tasks.
- A comprehensive summary of the existing surveys on visual transfer learning using auxiliary knowledge.
- An analysis of research gaps in the area of visual transfer learning using KGs which can be used as a basis for future research.

The remainder of this paper is structured as follows: Section 2 provides an overview of the methodology followed to conduct the survey. In Section 3 we introduce the term visual transfer learning. In addition, we outline different types of modeling structures of knowledge graphs such as directed labeled graphs, hypergraphs, and hyper-relational graphs. We explain the notion of features extractor, specifically referring to visual and semantic features. Further, we describe the term knowledge graph embedding and provide a brief categorization of KGE-Methods concerning different supervision and input types. Several joint training objectives suitable to combine semantic embeddings with visual embeddings are described. The main section, Section 4 introduces four different categories on how a KG can be combined with a DL pipeline:

1) *Knowledge Graph as a Reviewer* – where the KG is used for post-validation of a visual model;
2) *Knowledge Graph as a Trainee*, where the KG is embedded into $h_{s,v}$ using h_v as objective;
3) *Knowledge Graph as a Trainer*, where the KG with h_s is used as an objective to embedd images into $h_{v,s}$; and
4) *Knowledge Graph as a Peer*, where the KG with h_s is combined with h_v to suit as objective that embedds both the KG and images into h_h.

Since KGE-Methods have only recently entered the field of visual transfer learning, we also list related methods forming h_s based on other types of auxiliary knowledge in categories 2), 3), and 4). Other types of auxiliary knowledge are language descriptions or class attributes so that their semantic features extractor $f_s(\cdot)$ differs in the type of input, but not in its architecture. Furthermore, in Section 5 we provide an overview of generic KGs, several datasets and benchmarks using various types of auxiliary knowledge, like attributes, textual descriptions, or graphs. In Section 6 we summarize related surveys in the field of visual transfer learning and knowledge-based ML. Section 7 gives an outlook about challenges and open issues in the field of visual transfer learning using knowledge graphs. Finally, Section 8 provides a discussion and a conclusion as well as an outlook of future directions on the field.

2. Methodology

Our objective is to provide a comprehensive overview of how KGs can be used in combination with DL to solve visual transfer learning tasks. To ensure the quality of the outcome, we followed the process proposed by Petersen et. al [108,109] and conducted an initial search on five scholarly indexing services. We applied inclusion and exclusion criteria on our findings and extended them based on the snowballing approach [152].

2.1. *Research questions*

The combination of visual and semantic data seems to be a promising direction to build models that can cope with the diversity of the real world. However, some major challenges and questions arise when combining these modalities.

- **RQ1** – How can a knowledge graph be combined with a deep learning pipeline?
- **RQ2** – What are the properties of the respective combinations?
- **RQ3** – Which knowledge graphs already exist, that can be used as auxiliary knowledge?
- **RQ4** – What datasets exist, that can be used in the combination with auxiliary knowledge to evaluate visual transfer learning?

RQ1 and **RQ2** are answered in Section 4, where we categorize and discuss visual transfer learning approaches based on how the KG is combined with the DL pipeline. **RQ3** and **RQ4** are answered in Section 5, where we summarize available KGs, datasets, and benchmarks that will help to compare approaches of the field of visual transfer learning using KGs.

2.2. *Literature search*

To collect relevant literature, we define a search string using the following strategy:

- Extract major terms from research questions.
- Use synonyms and alternative terms.
- Combine using *OR* to include synonyms and alternative terms.
- Combine using *AND* to join the key terms.

As a result, the following major terms related to the concepts are derived: Knowledge Graph, Visual Transfer Learning, and connect them by a Boolean AND operation. Each term contains a set of keywords related to the respective concept, connected by a Boolean OR operation. Therefore, the initial search string was as follows: **((“Knowledge Graph” OR “Knowledge Graph Embedding” OR “Semantic Embedding”) AND (“Visual Transfer Learning” OR “Transfer Learning” OR “Zero-shot Learning” OR “Deep Learning” OR “Computer Vision”))**

For the primary search process we used five scholarly indexing services: Google Scholar,[1] IEEE Xplore,[2] ACM Digital Library,[3] Scopus,[4] and DBLP.[5]

[1] https://scholar.google.com
[2] https://ieeexplore.ieee.org
[3] https://dl.acm.org
[4] https://www.scopus.com
[5] https://dblp.uni-trier.de

2.3. Literature selection and quality assessment

After the literature search we included literature based on the following criteria:

- Studies using visual features.
- Studies using auxiliary knowledge.

Further, we excluded literature based on the following criteria:

- News articles.
- Non-English studies.
- Non-public available studies.
- Duplicate studies.

We reduced the amount of 16,200 studies after applying the inclusion and exclusion criteria on title and abstract to 17 relevant surveys and 164 studies (1.12%) During full-text reading, it became obvious that further articles should be removed as they were not in the scope based on the inclusion and exclusion criteria. The remaining articles (106) were used to conduct backward snowball sampling [152], which led to 22 additional studies. On the set of 128 primary studies we conducted a quality assessment on the following questions:

- Does the study provide a solid assessment?
- Are the results plausible?

Thus, we were able to reduce the number of studies to 124. These studies provide the basis for the survey and serve to answer the formulated research questions.

3. Background

This section briefly introduces the term visual transfer learning, describes the fundamentals of KGs, feature extractors, knowledge graph embeddings, and joint training objectives in the context of this survey.

3.1. Visual transfer learning

Visual transfer learning is presented in [118] as follows: *Given a source domain D_S with input data X_S, a corresponding source task T_S with labels Y_S, as well as a target domain D_T with input data X_T and a target task T_T with labels Y_T, the objective of visual transfer learning is to learn the target conditional probability distribution $P_T(Y_T|X_T)$ with the information gained from D_S and T_S where $D_S \neq D_T$ or $T_S \neq T_T$.*

Zero-shot learning is a visual transfer learning task with labeled source domain data and unlabeled target domain data. Zero-shot learning aims to extract implicit knowledge of the classes in the source domain task T_S and transfers this knowledge to unknown classes of the target domain task T_T [103]. If zero-shot learning has access to an additional set of labeled target data X_T, the task is called few-shot learning.

Domain generalization is a visual transfer learning task with access to labeled source domain data and unlabeled target domain data. Domain generalization aims to extract implicit knowledge of the source domain D_S and transfer this knowledge to an unknown target domain D_T [12,95]. If domain generalization has access to an additional set of labeled target data X_T, the task is called domain adaptation.

3.2. Knowledge graph

Knowledge is the awareness, understanding, or information for a phenomenon or a subject that has been obtained by observations or study.[6] It can be either implicit or explicit and stored and represented in different ways. Explicit

[6] https://dictionary.cambridge.org/dictionary/english/knowledge

knowledge is the type of knowledge that can be easily interpreted, organized, managed, and transmitted to others. Implicit knowledge is the form of knowledge that is gathered through observations and activities of everyday life. Using various modeling techniques, complex explicit knowledge can be formally represented in KGs. On the other hand, a common method for gathering implicit knowledge is to use feature extraction methods, that learn latent knowledge representations, e.g. visual or semantic embeddings, from observations [47].

There exist many ways for expressing, representing, and storing knowledge. In this survey, we focus on KGs, a structured representation of facts, consisting of entities, relationships, and semantic descriptions. A comprehensive definition is given by the authors of [58] where a KG is defined as *a graph of data with the objective of accumulating and conveying real-world knowledge, where entities are represented by nodes and relationships between entities are represented by edges*. Knowledge can be expressed in a factual triple in the form of (head, relation, tail). In its most basic form, we see a KG as a set of triples $G = H, R, T$, where H is a set of entities, $T \subseteq E \times L$, is a set of entities and literal values and R, set of relationships which connects H and R.

A graph model is a model which structures the data, including its schema and/or instances in form of graphs, and the data manipulation is realized by graph-based operations and adequate integrity constraints [3]. Each graph model has its formal definition based on the mathematical foundation, which can vary according to different characteristics, for instance, directed vs undirected, labeled vs unlabeled, etc. The most basic model is composed of labeled nodes and edges, easy to comprehend but inappropriate to encapsulate multidimensional information. Other graph models allow for the representation of information utilizing complex relationships in the form of hypernodes or hyperedges. In the following, we discuss three common graph models that are used in practice to represent data graphs.

Directed labeled graphs: A directed labeled graph is comprised of a set of nodes and a set of edges connecting those nodes, labeled based on a specific vocabulary [3].

The direction of the edge of two paired nodes is important, which clearly distinguishes between the start node and the end node. This intuitively enables the organization of information via the utilization of binary relationships.

Hypergraphs: Hypergraphs extend the definition of binary edges by allowing the modeling of multiple and complex relationships [3].

On the other hand, hypernodes modularize the notion of node, by allowing nesting graphs inside nodes. In addition, the notion of a hyperedge enables the definition of n-ary relations between different concepts.

Hyper-relational graphs: A hyper-relational graph is also a labeled directed multigraph where each node and edge might have several associated key-value pairs [4].

Internally, nodes and edges are annotated according to a chosen vocabulary and have unique identifiers, making them a flexible and powerful form of modeling for graph analysis with weighted edges.

Table 1 illustrates the three graph models mentioned above with some corresponding examples. A KG can be based on any such graph model utilizing nodes and edges as a fundamental modeling form.

3.3. Feature extractor

A feature extractor is a transformation function from higher dimensional into lower dimensional vector space, including a vast variety of dimensionality reduction methods [11,147].

Since it has been shown that most downstream tasks can be solved better on a reduced dimensionality, feature extractors are also a fundamental building block of modern systems working on visual and semantic data.

However, more and more conventional feature extraction methods have been replaced with DNNs. A DNN is an artificial *neural network* (NN) with multiple layers between the input and output layers, having the ability to automatically extract lower dimensional features from the input data [57,69].

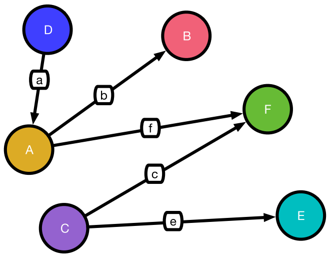
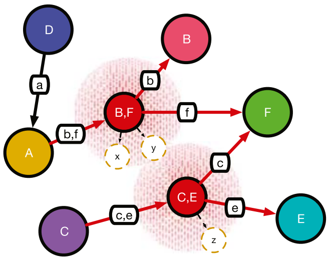
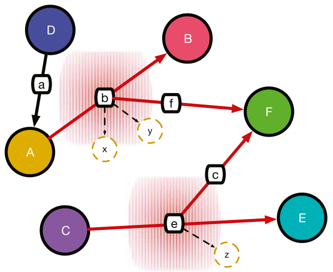
As depicted in Fig. 1, a DNN can be decoupled in a feature extractor $f(\cdot)$, with its embedding space $\vec{h}$ and a prediction task $g(\cdot)$, expressing the function

$$\hat{\vec{y}} = g\big(f(\vec{x})\big), \quad \text{with } f(x) = \vec{h}. \tag{1}$$

There are different architectures of DNNs, but they always consist of the same components: neurons, synapses, weights, biases, and functions [47]. The most common architectures that build a DNN are *multilayer perceptrons*

Table 1

Various graph models. Three common graph models used as underlying structure for knowledge representation in KGs: 1) directed labeled graphs; 2) hypergraphs; and 3) hyper-relational graphs

	Directed labeled graphs	Hypergraphs	Hyper-relational graphs
Nodes and Literals	– Real-world and abstract entities – Entity’s attribute value	– Real-world and abstract entities – Entity’s attribute value	– Real-world and abstract entities – Entity’s attribute value
Relationships	– Binary relations between entities – Relations between an entity and its attribute’s values	– Binary relations between entities – Relations between an entity and its attribute’s values – Many-to-many relations between entities (Hyperedge)	– Binary relations between entities – Relations between an entity and its attribute’s values – Additional information encoded in relationship (Hyper-relation)
Semantics	Connect two nodes	Connect an arbitrary set of nodes	Connect two nodes with additional contextual information
Example		 D B a b B,F f F b,f A x y c C,E c,e C e E z	 D B a b f F y x A c C e E z

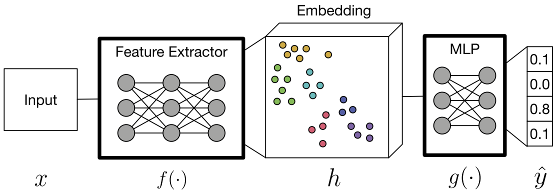


Fig. 1. A DNN that takes $\vec{x}$ as input and predicts $\hat{\vec{y}}$ can be decoupled into a feature extractor $f(\cdot)$ with its embedding space $\vec{h}$ and a prediction task $g(\cdot)$.

(MLP), *convolutional neural networks* (CNN), *recurrent neural networks* (RNN), and *transformer models*. Each architecture has its advantages and is therefore preferred for a particular type of input data and particular task [47].

Whereas, DNNs are usually trained end-to-end resulting in a task-dependent embedding space $\vec{h}$, more recently, attempts have been made to independently pre-train the feature extractor that it can be applied to several visual transfer learning and downstream tasks [22].

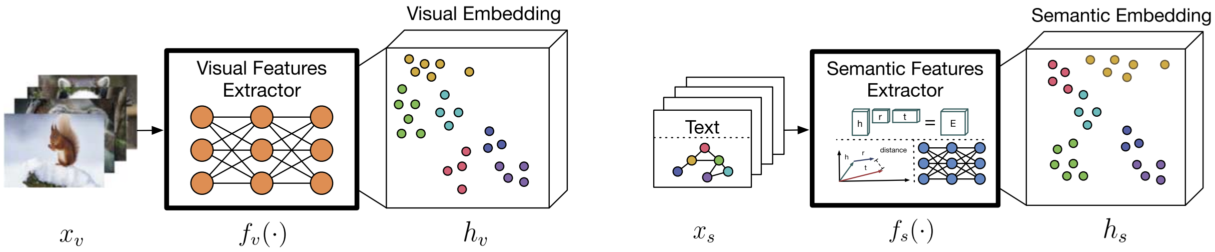
Visual features extractor: A visual features extractor $f_v(\cdot)$, shown in Fig. 2(a), is a transformation function that transform visual input data $\vec{x}_v$ from an higher dimensional image space into a lower dimensional visual embedding space $\vec{h}_v$.

A formal definition is given by

$$\vec{h}_v = f_v(\vec{x}_v), \tag{2}$$

where the final dimensionality of $\vec{h}_v$ is determined by the architecture.

Whereas early approaches used traditional visual features extractors as *scale-invariant feature transform* (SIFT) [87] or *histogram of oriented gradients* (HOG) [25], modern CV methods use almost only DNN-based approaches. A common method to obtain a general DNN-based visual feature extractor is to pre-train a DNN on a large image dataset, such that the DNN automatically learns to extract valuable features out of the images.



(a) Visual features extractor

(b) Semantic features extractor

Fig. 2. Feature extractors transform input data into embedding space: (a) a visual features extractor transforms visual input data, i.e. images, into visual embedding space; and (b) a semantic features extractor transforms semantic input data, e.g. text or graphs, into semantic embedding space.

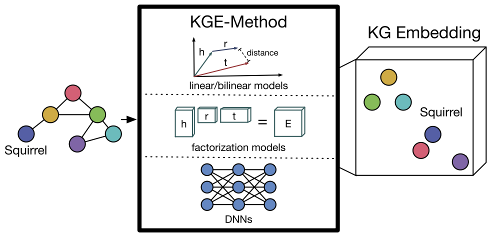


Fig. 3. A KGE-method transforms a KG into a knowledge graph embedding h_s.

Semantic features extractor: A semantic features extractor $f_s(\cdot)$, shown in Fig. 2(b), is a transformation function that transform semantic input data $\vec{x}_s$ from an higher dimensional image space into a lower dimensional semantic embedding space $\vec{h}_s$.

A formal definition is given by

$$\vec{h}_s = f_s(\vec{x}_s), \tag{3}$$

where the final dimensionality of $\vec{h}_s$ is determined by the architecture.

The term semantic data is here used for both, unstructured data from language and structured data from a KG. Although the input data structure differs in its original format, the output of the semantic features extractor is always a low dimensional and vector-based semantic embedding space. This similarity enables a seamless transfer from hybrid approaches of vision and language to hybrid approaches of vision and KGs.

3.4. Knowledge graph embedding

A knowledge graph embedding h_s is a representation of a KG in vector space, where close relationships between entities in a KG are reflected by local neighborhoods in h_s. h_s is generated by a *knowledge graph embedding method* (KGE-Method), which maps the entities and relations of a KG into low-dimensional vectors, while capturing their semantic meanings and relations [145]. Therefore, a KGE-Method is a special case of the semantic features extractors $f_s(\cdot)$ that works on graph data.

In Fig. 3, the general pipeline of KGE-Methods which transform a KG into h_s is illustrated.

3.4.1. KGE-methods – learning mode

Originally, KGE-Methods were developed to solve graph-based tasks such as node classification or link prediction. However, there is an increasing interest to apply KGE-Methods for visual tasks, such as classification, detection, or segmentation. We briefly categorize KGE-Methods therefore into unsupervised and supervised KGE-Methods, as Chami et al. [15] recently proposed for graph embedding algorithms.

Unsupervised KGE-methods: Unsupervised KGE-Methods form h_s based on the inherent graph structure and the node features, without considering additional task-specific labels for the graph or its nodes. An overview about unsupervised KGE-Methods is given by Ji et al. [62], who categorized KGE-Methods based on their *representation space* (vector, matrix, and tensor space), the *scoring function* (distance-based, similarity-based), the *encoding model* (linear/bilinear models, factorization models, neural networks), and the *auxiliary information* (text descriptions, type constraints).

Supervised KGE-methods: In contrast, supervised KGE-Methods learn h_s to best predict node or graph labels. Forming h_s by using task-specific labels for the node features, h_s can be optimized for a particular task while retaining the full expressivity of the graph. The most common supervised KGE-Methods are *graph neural networks* (GNNs) [49]. GNNs are extensions of standard DNNs that can directly work on a graph structure as provided by a KG. For scalability reasons and to overcome challenges that arise from graph irregularities various adaptations have emerged, such as *graph convolutional networks* (GCN) [71] or *graph attention networks* (GAT) [136]. Furthermore, non-Euclidean graph convolutional methods, such as *hyperbolic graph convolutional neural networks* (HGCN) [16] are used to deal with a hierarchical structure of the input data.

3.4.2. *KGE-methods – input type*

The majority of existing KGE-Methods only work on directed labeled graphs, expecting binary relations in a tripled-based format. However, as shown in Section 3.2, a basic triplet representation oversimplifies the complex nature of the information that can be stored in hypergraphs and hyper-relational graphs [116]. A hypergraph or hypher-relational graph can be transformed into directed labeled graphs, either by *reification* [35], that converts the graphs into binary-relation graphs, by creating additional triplets from a hyper-relational fact or by the *star-to-clique* [151] technique, that converts a tuple defined on k entities into $\binom{k}{2}$ tuples. However, these conversions lead to suboptimal and incomplete models as well as information loss. They only convert a set of key-value pairs, that are unaware of the triplet structure [35,116]. To preserve the whole expressivity of the KG, a set of new KGE-Methods are developed to directly operate on hypergraphs and hyper-relational graphs. Some of the methods that deal with hypergraphs are HEBE [52], HGE [157], Hyper2vec [59], HNN [36], HCN [154], DHNE [134], HHNE [9], Hyper-SAGNN [164], HypE [35] and methods that embedd hyper-relational graphs are for instance m-TransH [151], HSimple [35], RAE [163], GETD [84], TuckER [7], NaLP [51], HINGE [116], StarE [40].

3.5. *Training objectives for joint embeddings*

Since visual and semantic information can be encoded in a vector-based embedding space forming h_v and h_s, there are several training objectives to learn a joint embedding. The objectives and also the DNNs are optimized mainly using *stochastic gradient descent* (SGD) or its derivatives. SGD minimizes an objective, that measures how far apart the ground truth from the predicted probability distribution or value is. The most common principle to derive specific objectives that are good estimators for different models is the maximum likelihood principle. Any of these objectives can be seen as a cross entropy between the empirical distribution defined by the training set and the probability distribution defined by model [47]. Here we present some of the basic objectives used in visual transfer learning using KG, which can be augmented with additional regularization terms or hyperparameters. Although work [13,73] showed that the objectives have a smaller impact on the learned DNN than suspected, there are configurations of visual and semantic embedding space that only allow certain objectives to be applied. We define $\vec{l} \in \mathbb{R}^K$ as the network's output (logits) vector, and $\vec{t} \in 0, 1^K$ as the one-hot encoded vector of targets, where $\|t\|_1 = 1$. We refer to visual data as x_v and semantic data as x_s, and equally to visual embedding as h_v and semantic embedding as h_s.

3.5.1. *Pointwise objectives*

Softmax cross-entropy (CE) [14]: CE is the most common objective to learn multi-class classification tasks. The softmax represents a probability distribution over a discrete variable with K possible values, i.e. classes. CE learns the DNN end-to-end by comparing the logits $\vec{l}$ with the target vector $\vec{t}$ and is given by

$$L_{\mathrm{CE}}(\vec{l}, \vec{t}) = -\sum_{k=1}^{K} t_k \log\left(\frac{\exp(l_k)}{\sum_{j=1}^{K} \exp(l_j)}\right) \tag{4}$$

$$= -\sum_{k=1}^{K} t_k l_k + \log \sum_{k=1}^{K} \exp(l_k) \tag{5}$$

Mean squared error (MSE): MSE is the most intuitive way of attracting two vectors and is given by

$$L_{\text{MSE}} = \frac{1}{K} \sum_{k=1}^{K} \left\| (\vec{h}_{s,k} - \vec{h}_{v,k}) \right\|^2. \tag{6}$$

The MSE loss calculates the Euclidean distance and maps a training image $x_{v,k}$ and its visual feature vector $h_{v,k)}$ to a semantic embedding vector $h_{s,k}$, corresponding to the same class k [128].

However, using the Euclidian distance as a metric fails in high-dimensional space [89]. An alternative metric in high dimensions is the cosine distance, which is given by $\text{sim}(\vec{u}, \vec{v}) = \vec{u}^\top \vec{v} / \|\vec{u}\| \|\vec{v}\|$.

3.5.2. *Pairwise objectives*

Pairwise objectives [53] always rely on the information of positive and negative samples. They have the goal to pull positive visual embedding vectors $\vec{h}_{v,p}$ to its corresponding semantic embedding anchor vector $\vec{h}_{s,a}$ and push negatives $\vec{h}_{v,n}$ away [37].

Triplet and hinge rank loss [143]: The triplet and hinge rank loss requires an explicit negative sampling. It uses a margin α as a regularization term and it is given by

$$L_{\text{tri}} = \sum_{n \neq p} \max\left[0, \alpha - \text{sim}(\vec{h}_{s,a}, \vec{h}_{v,p}) + \text{sim}(\vec{h}_{s,a}), \vec{h}_{v,n}\right]. \tag{7}$$

Contrastive loss: The contrastive loss extends the triplet loss by a version of the softmax and handles multiple positives and negatives at a time and is given by

$$L_{\text{con}} = -\log \frac{\exp(\text{sim}(\vec{h}_{s,a}, \vec{h}_{v,p})/\tau)}{\sum_{n=1}^{2N} \mathbb{1}_{n \neq a} \exp(\text{sim}(\vec{h}_{s,a}, \vec{h}_{v,n})/\tau)} \tag{8}$$

where, $\mathbb{1}_{n \neq a} \in \{0, 1\}$ is an indicator function that returns 1 iff $n \neq a$, and $\tau > 0$ denotes a temperature parameter.

4. Visual transfer learning using knowledge graphs

Visual transfer learning using knowledge graphs has proven to be particularly advantageous compared to approaches without auxiliary knowledge [128,148]. Since auxiliary knowledge mitigates the sole dependence on data distribution, it leads to models that are better generalized and thus more robust and applicable to new domains [78]. Having various kinds of auxiliary knowledge, a KG can serve as a universal knowledge representation. KGs encode the classes either hierarchically, organized in superclasses, or flat, using relationships to other objects or other classes. Section 3.2 presents three distinct modeling structures with different levels of expressiveness and Section 3.4 introduces relevant embedding methods. All approaches that use a KG in combination with a DNN use the KG to implement some prior assumptions in the data-driven DL pipeline. A prior assumption induced by the KG is the definition of relationships between objects/classes so that objects/classes can borrow statistical strength from other related objects/classes in the graph. These priors give the CV process a structure that allows making better predictions even when visual data is sparse or erroneous. However, there are several ways the auxiliary knowledge of a KG can be induced into a DNN.

Referring to **RQ1**, this section provides a categorization of visual transfer learning approaches that combine KGs with the DL pipeline.

As shown in Fig. 4, we categorize the field of visual transfer learning using knowledge graphs into:

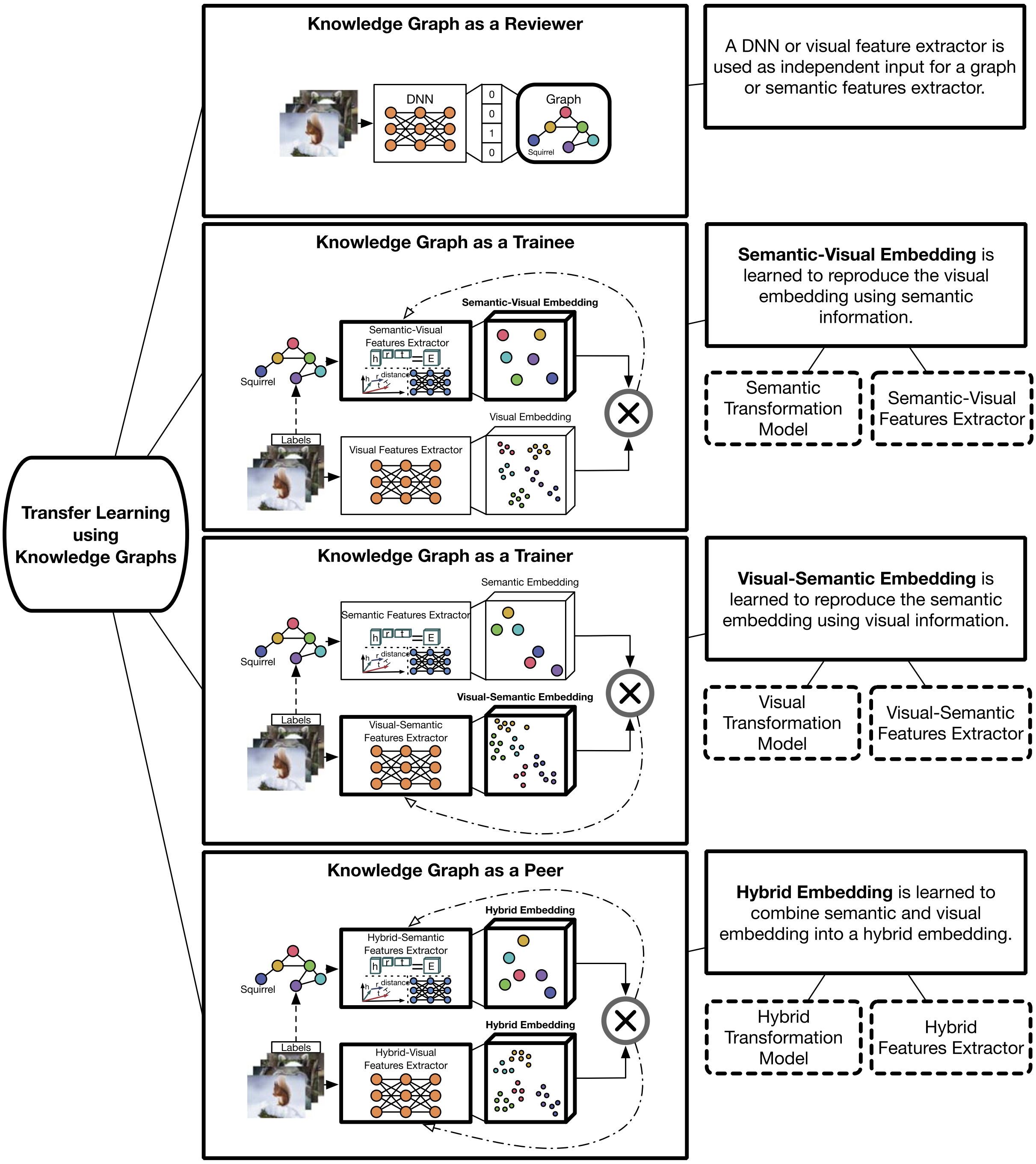

Transfer Learning using Knowledge Graphs
Knowledge Graph as a Reviewer
DNN
0
0
1
0
Graph
Squirrel
A DNN or visual feature extractor is used as independent input for a graph or semantic features extractor.
Knowledge Graph as a Trainee
Semantic-Visual Embedding
Semantic-Visual Features Extractor
h
r
t
E
distance
Squirrel
Labels
Visual Embedding
Visual Features Extractor
Semantic-Visual Embedding is learned to reproduce the visual embedding using semantic information.
Semantic Transformation Model
Semantic-Visual Features Extractor
Knowledge Graph as a Trainer
Semantic Embedding
Semantic Features Extractor
Visual-Semantic Embedding
Visual-Semantic Features Extractor
Visual-Semantic Embedding is learned to reproduce the semantic embedding using visual information.
Visual Transformation Model
Visual-Semantic Features Extractor
Knowledge Graph as a Peer
Hybrid Embedding
Hybrid-Semantic Features Extractor
Hybrid Embedding
Hybrid-Visual Features Extractor
Hybrid Embedding is learned to combine semantic and visual embedding into a hybrid embedding.
Hybrid Transformation Model
Hybrid Features Extractor

Fig. 4. Visual transfer learning using KGs according to the role of the KG are split in four categories: 1) *knowledge graph as a reviewer*; 2) *knowledge graph as a trainee*; 3) *knowledge graph as a trainer*; and 4) *knowledge graph as a peer*.

1) *Knowledge Graph as a Reviewer* – where the KG is used for post-validation of a visual model;

2) *Knowledge Graph as a Trainee*, where a semantic-visual embedding $h_{s,v}$ is learned using a visual embedding h_v as objective;

3) *Knowledge Graph as a Trainer*, a visual-semantic embedding $h_{v,s}$ is learned using a semantic embedding h_s

Table 2

Categories and their tasks: task transfer refers to the category zero and few-shot learning, domain transfer refers to the category domain generalization and adaptation, and other relates to object classification, object detection, and object segmentation on source task and domain only. Note: all approaches using related types of auxiliary knowledge are highlighted in bold

Category	Sub-category	Task transfer	Domain transfer	Other
Knowledge Graph as a Reviewer		[27,29,76,88,115]	[42,46]	[23,63,82,85,90,99,121]
Knowledge Graph as a Trainee	Semantic-Visual Transformation Model	**[114,162]**		
	Semantic-Visual Features Extractor	[21,41,45,67,106,148], **[80,141,169]**		[24]
Knowledge Graph as a Trainer	Visual-Semantic Transformation Model	[2], **[37,72,98,102,128,166]**		**[93]**
	Visual-Semantic Features Extractor	[61]	[94], **[110]**	**[65]**
Knowledge Graph as a Peer	Hybrid Transformation Model	[117,155,167], **[6,17,38,64,133]**	**[155]**	**[68,81,132,158]**
	Hybrid Features Extractor	[96]		**[165]**

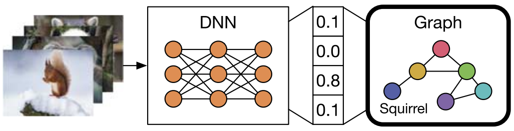


Fig. 5. Approaches from the category *knowledge graph as a reviewer* use the KG for post-validation of a pre-trained DNN or its intermediate feature layers.

as objective; and

4) *Knowledge Graph as a Peer*, where a hybrid-embedding h_h is learned using a combination of semantic embedding h_s and a visual embedding h_v as objective.

Since KGE-Methods have only recently entered the field of visual transfer learning, we also list related methods forming h_s based on other types of auxiliary knowledge in categories 2), 3), and 4). Other types of auxiliary knowledge are language descriptions or class attributes, so that their semantic features extractor $f_s(\cdot)$ differs in the type of input, but not in its architecture, as described in Section 3.3.

Regarding **RQ2**, we describe the categories and their approaches in detail and discuss their field of application and their properties. A summary of all approaches and their respective transfer learning task is given in Table 2.

4.1. Knowledge graph as a reviewer

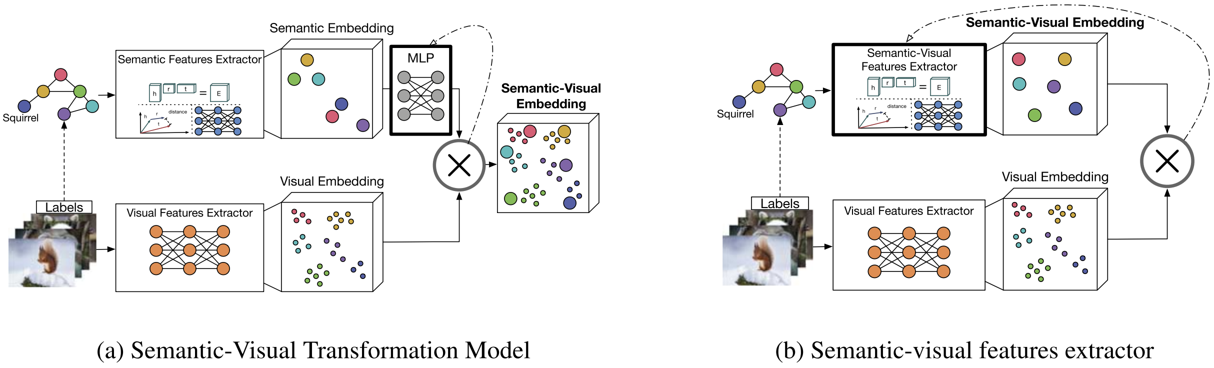
Approaches of the category *Knowledge Graph as a Reviewer* arrange the visual model and the KG in a sequential order, as depicted in Fig. 5. The visual output of a pre-trained DNN or its intermediate feature layers suit as an input to a graph or graph-based network. Unlike the other categories, the KG as a reviewer does not learn a joint embedding space, instead, it uses the KG or its h_s to reason over the independent output of a visual model h_v.

Most of the approaches map the output of a visual features extractor $f_v(\cdot)$ on the corresponding input nodes in a hierarchical graph, to enrich the output with inter-class relationships. Lampert et al. [76] train a *support vector machine* (SVM) on SIFT features to predict binary *animals with attributes* (AwA) dataset attributes. These class attributes are fed into a hierarchical graph-based network to predict unknown classes for a zero-shot learning task. Salakhutdinov et al. [121] introduce a hierarchical Bayesian classification model [123] that learns a tree structure of class and super-class relationships. They use their learned graph on top of an SVM, which classifies HOG features of images. They show that their method using a learned graph outperforms a method using a fixed graph based on

WordNet[7] [92] and other approaches without hierarchical graph information. Deng et al. [29] proposed the *DARTS* algorithm for zero-shot learning. They pre-train an SVM on SIFT features of the ImageNet [28] dataset and map its classification output to WordNet with a reward and an accuracy to maximize the information gain. Ordonez et al. [99] extend the approach to output human-understandable entry categories for images. They enrich the output of an SVM-based image classification model with information from a text-based n-gram language model by mapping both sources to the corresponding node in the WordNet graph. Rohrbach et al. [115] present *propagated semantic transfer* (PST). They use WordNet and attribute vectors from the AwA dataset to perform classification on few-shot learning classes of ImageNet. PST exploits similarities in visual embeddings of known classes encoded by an SVM learning a *k-Nearest Neighbor* (kNN) graph that helps to find relationships to new classes. Deng et al. [27] propose to use a *hierarchy and exclusion* (HEX) graph that exploits hierarchical class relationships of the output of a visual model. HEX graphs allow flexible specification of relations between labels applied to the same object. To build the graph, they use the hierarchical structure of WordNet extended with additional specifications and relations to objects, such as mutual exclusion (e.g., an object cannot be a dog and a cat), overlap (e.g., a husky can be a puppy and vice versa), and subsumption (e.g., all huskies are dogs). In addition, they proposed a probabilistic classification model that exploits their HEX graphs and evaluated their approach on ImageNet, in object classification and zero-shot learning. Gebru et al. [42] use WordNet attributes to improve fine-grained object classification on the task of domain generalization with the Office-31 [120] and the large-scale Car dataset [43]. Source and target domain images are fed through a pipeline with two identical CNNs and a classification layer that classifies both the fine-grained classes and the different attribute types. The Kullback–Leibler divergence is used to compare the predicted label distributions. Lee et al. [79] propose a *graph gated neural network* (GGNN) that incorporates a structured KG based on WordNet and learned edge weights to improve zero-shot learning. First, an NN is learned that combines the GloVe [107] language embeddings of the class labels and the pre-trained visual feature vectors of the images as input to the GGNN. Second, the GGNN learns to propagate the information through the KG and outputs a final probability for each node.

Instead of using hierarchical graphs of WordNet and class attributes only, other approaches make use of flat object or class relationships. Their graph consists of specific real-world configurations of objects and their appearance. Marino et al. [90] improves fine-grained image classification by creating a KG using the most common object-attribute and object-object relationships of the Visual Genome [74] dataset and higher-level semantics from WordNet. The output of a pre-trained, faster R-CNN [113] object detector is fed into a *graph search neural network* (GSNN) which reasons about relationships of the detected objects. The final prediction is a combination of the GSNN output, the visual embedding, and the detections of the faster R-CNN. Chen et al. [23] propose an object detection post-processing that connects a local and a global module via an attention mechanism. The local module is based on a convolutional *gated recurrent unit* (GRU) and builds spatial memory of previously detected objects using the class label and its visual embedding. The global graph-reasoning module consists of two paths, a spatial path that uses a region graph to connect far detected classes, and a semantic path which uses a KG, based on ADE20K [168] and Visual Genome, to connect classes with semantically related classes. Jiang et al. [63] extend [23] with *hybrid knowledge routed modules* (HKRM) allowing them to be applied on the intermediate feature representation directly to check the compatibility of auxiliary knowledge with visual evidence in each image. HKRM can be divided into an explicit knowledge module and an implicit knowledge module, whereas the former contains external knowledge such as shared attributes, co-occurrence, and relationships, and the latter is built without explicit definitions and forms a region-to-region graph with constraints over objects, as spatial knowledge such as layout, size, overlap. Liu et al. [85] improve object detection by feeding the final object detections into a GCN which is based on object relationships and learned from MSCOCO dataset [83]. Gong et al. [46] propose a human parsing agent called "Graphonomy" that learns a knowledge graph on a conventional parsing network. It consists of an intra-graph reasoning module in form of a GCN whose structure uses semantic constraints from the human body to transfer knowledge within a dataset due to encoded relationships between nodes, and an inter-graph reasoning module, that uses handcrafted relations, a learnable matrix, feature similarities, and semantic similarities, to transfer semantic information between different datasets. Liang et al. [82] present a *symbolic graph reasoning* (SGR) layer

[7] https://wordnet.princeton.edu/


Semantic Embedding
Semantic Features Extractor
MLP
Semantic-Visual Embedding
Squirrel
Labels
Visual Features Extractor
Visual Embedding
Semantic-Visual Embedding
Semantic-Visual Features Extractor
Squirrel
Labels
Visual Features Extractor
Visual Embedding

(a) Semantic-Visual Transformation Model

(b) Semantic-visual features extractor

Fig. 6. Approaches that belong to the category *knowledge graph as a trainee* learn semantic visual embedding space supervised by a visual embedding. They either learn (a) a transformation function, e.g. MLP, on top of a pre-trained semantic embedding space or (b) a semantic-visual features extractor.

for semantic segmentation and image classification. It consists of a module that assigns the visual features of a pre-trained DNN to corresponding nodes of a KG. A graph reasoning over all previously defined nodes is performed, and a mapping from the symbolic graph information back to the visual feature space. Their graph is based on an object relation graph from Visual Genome and a hierarchical relation graph from WordNet.

Luo et al. [88] propose a context-aware zero-shot learning framework, where they use a KG to reason about visual feature vectors generated from an object detection model. By using inter-class relationships, they improve traditional zero-shot learning techniques on the Visual Genome dataset.

4.2. *Knowledge graph as a trainee*

Approaches that belong to this category combine the visual DNN with the auxiliary knowledge of a KG by learning a semantic-visual embedding $h_{s,v}$. Unlike the *Knowledge Graph as a Reviewer*, which uses the visual embedding h_v as input for the KG, approaches from the category *Knowledge Graph as a Trainee* use h_v as an objective to embedd the KG into $h_{s,v}$. Figure 6 illustrates a conceptual architecture of the knowledge graph as a trainee approach. To combine visual and semantic information, some approaches either learn a transformation function, e.g. MLP, on top of a semantic embedding space h_s, or apply supervised KGE-Methods to learn a semantic-visual features extractor $f_{s,v}(\cdot)$ directly.

4.2.1. *Semantic-visual transformation models*

As shown in Fig. 6(a), the pre-trained h_s is fixed over the whole training process, and an additional transformation function, e.g. MLP, is learned to transform h_s, into the semantic-visual embedding space $h_{s,v}$.

Related approaches using other auxiliary knowledge: Rochan et al. [114] used a fixed language embedding to define relationships between classes, that unknown classes in a zero-shot learning task can borrow their visual embeddings from a linear combination of known related classes. Zhang et al. [162] extends suggesting to use the visual space, instead of the semantic space, as the main embedding space, thus reducing the hubness problem that occurs in high dimensions.

4.2.2. *Semantic-visual features extractors*

As illustrated in Fig. 6(b) the semantic-visual features extractor $f_{s,v}(\cdot)$ learns to directly transform the KG into a semantic-visual embedding $h_{s,v}$ using the supervision of the visual embedding space h_v. As described in Section 3.4, $f_{s,v}(\cdot)$ is mostly implemented using a supervised KGE-Method.

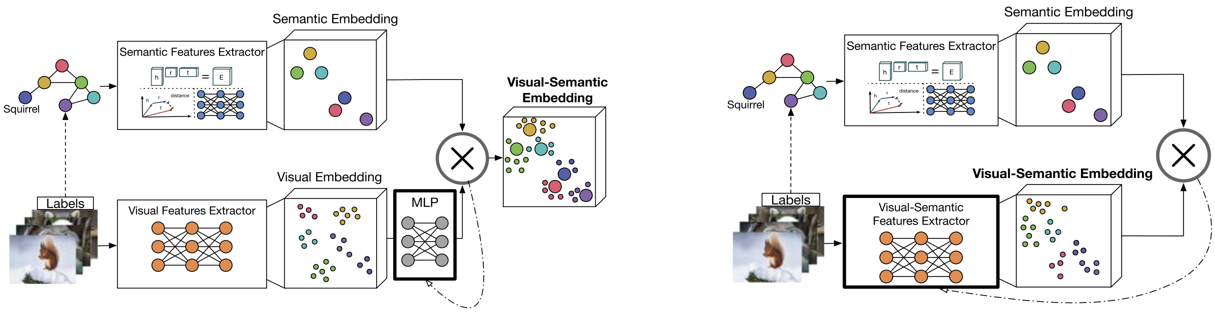
Wang et al [148] build a GCN on the structure of WordNet and optimize it to predict ImageNet pre-trained visual classifiers. Based on the learned relations in the GCN they are able to transform information to novel class nodes to perform zero-shot learning. A similar principle is used by Chen et al. [24] for multi-label image recognition. However, instead of using a hierarchical graph, the approach uses an object-relation graph which reflects the different relations between objects in a scene. They build their graph based on the occurrence probabilities of different objects in the MSCOCO dataset since some objects are more likely to occur together. Kampffmeyer et al. [67] claim that multi-layer GNN architectures, which are required to propagate knowledge to distant nodes in the graph, dilute the knowledge by performing extensive Laplacian smoothing at each layer and thereby consequently decrease performance. They propose a *dense graph propagation* (DGP) module with direct links among distant nodes to exploit the hierarchical graph structure of the KG. They tested their approach on zero-shot learning tasks as 21K ImageNet dataset and AWA2. Gao et al. [41] designed a *two-stream GCN* (TS-GCN) to perform *zero-shot action recognition* (ZSAR). Their GCN architectures are based on the ConceptNet 5.5 KG, which contains information from various knowledge bases such as WordNet and DBpedia. The first classifier branch uses the language embedding vectors of all classes as input for a GCN and then generates the classifiers for each action category. The second instance branch feeds video segments into a DNN and outputs object scores, which are combined with attribute vectors from the classifier branch using a post-processing GCN to form an attribute feature space. The final objective is then defined by a comparison of the attribute feature space and the output of the classifier branch. Peng et al. [106] propose a *knowledge transfer network* (KTN), which extends [148] with a vision-knowledge fusion model. This vision-knowledge fusion model is used to combine the final prediction output of the GCN with the output of a DNN, as they claim that semantic embeddings and visual embeddings are complementary and therefore cannot be combined with a single inner product. They pre-train their visual feature learning module using cosine similarity on image data, use a subgraph of WordNet for their knowledge transfer module, and language embeddings of the class labels as the initial state of the nodes of the GCN. Chen et al. [21] present the *knowledge graph transfer network* (KGTN). The knowledge graph transfer module incorporates a GGNN, which supports knowledge transfer of classes through a KG. To train GGNN, they fix the weights of a pre-trained visual features extractor and examine three different similarity metrics, such as inner product, cosine similarity, and person correlation coefficient, to compare the output of the DNN and the GGNN. They show that the accuracy of the model benefits from a reasoning process and the auxiliary knowledge from a KG.

Geng et al. [45] recently proposed Onto-ZSL, an ontology-enhanced zero-shot learning framework that can be applied either to image classification or knowledge graph completion. They build an inter-class relationship using an ontological schema, that comprises a label taxonomy from WordNet, textual descriptions, and attribute descriptions. Further, they address the data imbalance problem between seen and unseen images by leveraging a *generative adversarial network* (GAN) that produces synthesized visual feature vectors for unseen classes.

Related approaches using other auxiliary knowledge: Approaches using language models leverage GANs to imagine unseen categories from text descriptions and hence recognize novel classes with no examples being seen. GANs can be seen as a transformation function from text-based input to visual features, using the supervision of a visual model. Zhu et al. [169] propose GAZL, an approach that takes noisy text descriptions about unseen classes from Wikipedia and generates synthesized visual features for this class. Using textual input for unseen classes they learn a GAN that generates visual features similar to the pre-trained ones of the seen classes. Therefore, the zero-shot learning problem is transformed into a standard classification task and a classifier that can handle unseen classes can be trained using the synthesized image features for every unseen class. Li et al. [80] extended the approach by introducing LisGAN, a GAN that takes semantic descriptions and random noise to generate visual features for unseen classes. In addition, they deploy the average representation of all samples from an unseen class defining the soul sample of the class to reduce the noise in the predictions. Vyas et al. [141] propose LsrGAN, a generative model that leverages the semantic relationship between seen and unseen categories and explicitly performs knowledge transfer by incorporating a novel *semantic regularized loss* (SR-Loss). Knowing the inter-class relationships in the semantic space helps to impose the same relationship constraints among the generated visual features.

4.3. Knowledge graph as a trainer

Methods that belong to the category *Knowledge Graph as a Trainer* combine the visual output of a DNN with the auxiliary knowledge of a KG by learning a visual-semantic embedding $h_{v,s}$. Figure 7 illustrates a conceptual architecture of the knowledge graph as a trainer approach. The KG acts as a trainer and supervises the training of the DNN using h_s, rather than letting the DNN learn a h_v solely depending on the data distribution of the images. We refer to such an embedding of visual information learned under the supervision of a semantic embedding h_s


Semantic Embedding
Semantic Features Extractor
Squirrel
Visual-Semantic Embedding
Visual Embedding
Labels
Visual Features Extractor
MLP
Semantic Embedding
Semantic Features Extractor
Squirrel
Visual-Semantic Embedding
Labels
Visual-Semantic Features Extractor

(a) Visual-semantic transformation model

(b) Visual-semantic features extractor

Fig. 7. Approaches that belong to the category *knowledge graph as a trainer* learn visual semantic embedding space supervised by a semantic embedding. They either learn (a) a transformation function, e.g. MLP, on top of a pre-trained visual embedding space that suits as a transformation function or (b) a visual-semantic features extractor that learns the final embedding directly.

as a visual-semantic embedding $h_{v,s}$. To combine semantic and visual information, some approaches either learn a transformation function, e.g. MLP, on a pre-trained and fixed visual embedding h_v or learn a visual-semantic features extractor $f_{v,s}(\cdot)$ directly.

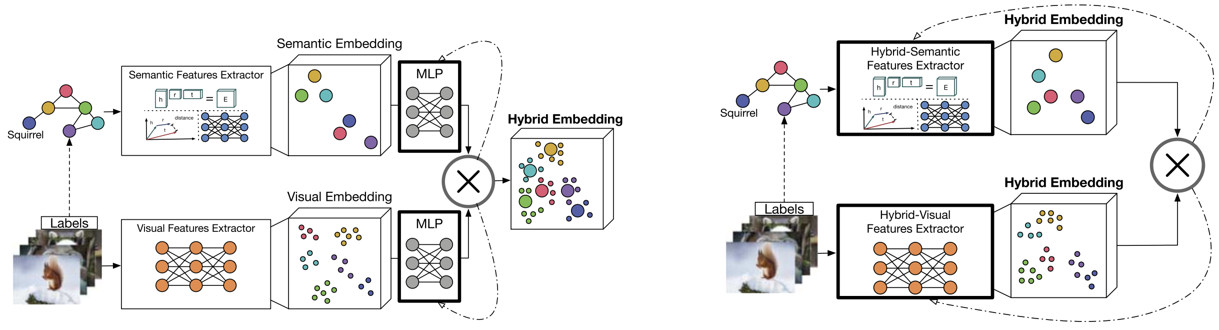
4.3.1. Visual-semantic transformation models

As shown in Fig. 7(a), the pre-trained h_v is fixed over the whole training process and an additional transformation function, e.g. MLP, is learned to transform h_v, into the visual-semantic embedding space $h_{v,s}$.

Akata et al. [2] refer to their semantic embedding space transformations as label embedding methods. They compared transformation functions from the visual embedding space to the attribute label embedding space, the hierarchy label embedding space, and the Word2Vec [91] label embedding space. Lonij et al. [86] approached the task of open-world visual recognition by using KGs. They learn h_s from a WordNet KG by using the *neural tensor layer* (NTL) [127] architecture and embedd the visual embedding generated by a pre-trained CNN into the same space using the hinge rank loss.

Related approaches using other auxiliary knowledge: One of the first approaches that use semantic embeddings with NNs is the work from Mitchell et al. [93]. They use language embeddings derived from text corpus statistics to generate neural activity pattern images. Instead of generating images from text, Palatucci et al. [102] learn a linear regression model to map neural activity patterns into language embedding space. Socher et al. [128] present a model for zero-shot learning that learns a transformation function between a visual embedding space, obtained by an unsupervised feature extraction method, and a semantic embedding space, based on a language model. The authors trained a 2-layer NN with the MSE loss to transform the visual embedding into the language embedding of 8 classes. Frome et al. [37] introduce the deep visual-semantic embedding model DeViSE that extends the approach from 8 known and 2 unknown classes to 1,000 known and 20,000 unknown classes. Therefore, they pre-train their visual features extractor using ImageNet and their semantic embedding vector using a skip-gram language model [91]. In contrast to Socher et al. [128] they learn a linear transformation function between the visual embedding space and the semantic embedding space using a combination of dot-product similarity and hinge rank loss since they claim that MSE distance fails in high dimensional space. Norouzi et al. [98] propose *convex combination of semantic embeddings* (ConSE). ConSE performs a convex combination of known classes in the semantic embedding space, weighted by their predicted output scores of the DNN, to predict unknown classes in a zero-shot learning task. Similarly, Zhang et al. [166] introduce the *semantic similarity embedding* (SSE), which models target data instances as a mixture of seen class proportions. They built a semantic space that each novel class could be represented as a probabilistic mixture of the projected source attribute vectors of the known classes.

Kodirov et al. [72] propose SAE a semantic autoencoder for zero-shot learning. It is learned by encoding pre-trained visual features of a CNN into a latent semantic space and then by decoding them back into visual space. The semantic space is based on class attributes for smaller datasets and on a word2vec language model for larger datasets.


Semantic Embedding
Semantic Features Extractor
Squirrel
MLP
Hybrid Embedding
Labels
Visual Embedding
Visual Features Extractor
MLP
Hybrid Embedding
Hybrid-Semantic Features Extractor
Squirrel
Hybrid Embedding
Labels
Hybrid-Visual Features Extractor

(a) Hybrid transformation model

(b) Hybrid features extractor

Fig. 8. Approaches that belong to the category *knowledge graph as a peer* learn hybrid embedding space as a combination of visual and semantic embedding space. They either learn (a) transformation functions, e.g. MLPs, on top of both pre-trained visual and semantic embedding spaces that suit as a transformation function or (b) hybrid features extractors that learn the final embedding directly.

They claim that their latent semantic embedding space can better handle the projection domain shift problem, i.e. the distribution shift between seen and unseen classes.

4.3.2. *Visual-semantic features extractors*

As illustrated in Fig. 7(b) the visual-semantic features extractor $f_{v,s}(\cdot)$ is learned to directly transform the images into a visual-semantic embedding $h_{v,s}$ using the supervision of the semantic embedding space h_s. As described in Section 3.4, h_s is mostly learned using an unsupervised KGE-Method and $f_{v,s}(\cdot)$ is implemented using a standard DNN.

Monka et. al [94] propose KG-NN, an approach that uses a KG and its h_s to train a visual DNN. Using a contrastive knowledge graph embedding loss in combination with h_s they learn a visual-semantic features extractor $f_{v,s}(\cdot)$. They test their approach on domain generalization and adaptation tasks for road sign recognition in Germany and China, as well as on mini-ImageNet and various derivatives. They show that their visual features extractor learned using the *Knowledge Graph as a Trainer* outperforms a conventional DNN trained with CE, the same DNN without additional information from the KG, and the same DNN using additional information from a pre-trained GloVe embedding in visual transfer learning tasks.

Jayathilaka et al. [61] proposed a framework named ViOCE that integrates ontology-based background knowledge in the form of n-ball class embeddings into a DNN-based vision architecture. The approach consists of two components – converting symbolic knowledge of an ontology into continuous space by learning n-ball embeddings that capture properties of subsumption and disjointness and guiding the training and inference of a vision model using the learned embeddings.

Related approaches using other auxiliary knowledge: Joulin et al. [65] demonstrate that feature extractors trained to predict words in image captions learn useful image representations. They convert the title, description, and hashtag metadata of images into a bag-of-words multi-label classification task and showed that pre-training a feature extractor to predict these labels learned representations which performed similarly to ImageNet-based pre-training on transfer tasks. Radford et al. [110] claim that state-of-the-art CV systems are restricted to predict a fixed set of predetermined object categories. Therefore, they propose to use a simple and general pre-training of their CNN with natural language supervision, i.e. predicting which caption goes with which image on a dataset of 400 million image-text pairs collected from the internet using the objective of Zhang et al. [165].

4.4. *Knowledge graph as a peer*

Approaches of the category *Knowledge Graph as a Peer* combine the visual DNN with the auxiliary knowledge of a KG by influencing both semantic and visual embedding. Unlike the previous categories, the idea of a hybrid embedding h_h is to fuse the visual embedding h_v and the semantic embedding h_s. Both semantic and visual data are then embedded into h_h. Figure 8 illustrates a conceptual architecture of the knowledge graph as a peer approach. The final hybrid embedding space is either a combination of pre-trained visual embedding h_v and semantic embedding h_s, using a transformation function, e.g. MLP, or a combination of hybrid-visual $f_{h,v}(\cdot)$ and hybrid-semantic features extractors $f_{h,s}(\cdot)$.

4.4.1. Hybrid transformation models

As shown in Fig. 8(a), pre-trained h_s and pre-trained h_v are fixed over the whole training process and an additional transformation functions, e.g. MLPs, are learned to transform h_s and h_v, into the hybrid embedding space h_h.

Zhao et al. [167] propose a joint model that combines an image stream and a concept stream via a joint loss function to preserve concept hierarchy as well as visual feature similarities. The concept stream is based on a language embedding with the hierarchical graph of WordNet and the image stream is a visual embedding from semantic segmentation DNN. They compare their approach against the standard CE-based approach and semantic embedding space transformations based on Word2Vec. Roy et al. [117] introduce a zero-shot learning model that takes advantage of the commonsense knowledge graph ConceptNet 5.5 to generate h_s of the class labels by using a GCN-based autoencoder. They enrich h_s with additional attributes and language embeddings, which is then compared with a pre-trained visual output of a DNN using a relation network [130].

Related approaches using other auxiliary knowledge: Yang et al. [155] propose a two-sided NN to learn a combination of a pre-trained visual embedding and a semantic embedding of attributes and word vectors based on image descriptions to perform zero-shot learning and domain generalization. To train their NN they use a Euclidean loss for regression and a hinge rank loss for classification. Fu et al. [38] try to reduce the bias of semantic embedding spaces, by proposing a transductive multi-view embedding framework that aligns novel features with the semantic embedding space for zero-shot learning. The framework first transforms the semantic embedding space into a joint embedding space using the unlabeled target data with a multi-view *canonical correlation analysis* (CCA) to alleviate the projection domain shift problem. And Second, a heterogeneous multi-view hypergraph label propagation method is used to perform zero-shot learning in the transductive embedding space, which combines additional semantic knowledge in the form of attributes and word vectors from related classes. Ba et al. [6] introduce a flexible zero-shot learning model that learns to predict unseen image classes using a language embedding. Therefore, they add two separate MLPs on top of the visual embedding and the semantic embedding and train them using the binary-CE loss, the hinge loss, and the Euclidean distance loss. Karpathy et al. [68] learn a model that generates language descriptions for detected objects in an image. Their objective aligns the output of a pre-trained CNN applied to image regions, and the output of a bidirectional RNN applied to sentences. Changpinyo et al. [17] use a set of "phantom" object classes whose coordinates live in both the semantic space and the model space. To align the two spaces, they view the coordinates in the visual embedding as the projection of the vertices on the graph from the semantic embedding. To compute low-dimensional Euclidean space embeddings from the weighted graph they propose to use the algorithm of Laplacian eigenmaps, mapping semantic and visual embedding into a common space defined by the mixture of seen classes proportions. Tsai et al. [133] propose the approach ReViSE that learns an unsupervised joint embedding of semantic and visual features to enable zero-shot learning. As external knowledge, they experiment with three different embedding methods for their attributes, human-annotated attributes [77], Word2Vec attributes, and GloVe attributes. Tang et al. [132] propose the *large scale detection through adaptation* (LSDA) framework to improve object detectors with image classification DNNs, hence without requiring expensive bounding box annotations. LSDA defines visual similarity as the distance between pre-trained visual embedding vectors and semantic similarity as the distance between pre-trained language embedding vectors of the labels. Jiang et al. [64] introduce their *transferable contrastive network* (TCN) explicitly transfers knowledge from the source classes to the target classes, to counteract the overfitting problem on source classes. To compute the similarities between classes in the hybrid embedding space, they design a contrastive network that automatically judges how well the embedding vector is consistent with a specific class. Li et al. [81] propose a multi-layer transformer [135] model as DNN, which uses object tags detected in images as anchor points to learn a joint embedding of the detected objects and the language tags, instead of simply concatenating visual embedding and semantic embedding. Yu et al. [158] propose a knowledge-enhanced approach, ERNIE-ViL, to learn joint representations of vision and language using a transformer model as DNN. ERNIE-ViL tries to construct the detailed semantic connections across vision and language while constructing a scene graph parsed from sentences and type prediction tasks, i.e., object prediction, attribute prediction, and relationship prediction in the pre-training phase.

4.4.2. *Hybrid features extractors*

As depicted in Fig. 8(b), hybrid-semantic $f_{h,s}(\cdot)$ and hybrid-visual $f_{h,v}(\cdot)$ features extractors are learned to directly transform KG and images into a common hybrid embedding h_h. As described in Section 3.4, $f_{h,s}(\cdot)$ is usually implemented using a supervised KGE-Method and $f_{h,v}(\cdot)$ using a standard DNN.

Recently, Naeem et. al [97] proposed a method to perform zero-shot image classification using hybrid features extractors. An ImageNet pre-trained DNN is used for the visual features extractor and a GCN in the *compositional graph embedding* (CGE) setting is used for the semantic features extractor. However, they learn a joint embedding function that can influence the weights of the DNN as well as the weights from the GCN. Interestingly, they compare their model against a similar version of their model, but with a fixed visual features extractor where the KG just acts as a trainee (see Section 4.2). They use that version for comparison with related approaches, stating that all other methods are based on fixed visual features extractors. Moreover, they show that a hybrid approach with an adaptive visual features extractor performs better than the other.

Related approaches using other auxiliary knowledge: Zhang et al. [165] use two contrastive pre-training objectives, contrasting semantic embedding to visual embedding, and vice versa, on the special domain of medical imaging to learn a joint feature extractor. Instead of previous works that learn transformation functions on top of fixed image trained visual features extractors they directly supervise the training of the CNNs with language embedding information. To train their DNN they use text-image paired data.

5. Visual transfer learning datasets and benchmarks

Building expressive knowledge graphs from scratch can be a quite challenging task. Concerning **RQ3**, this section provides an overview of standard and large-scale KGs that can be used as auxiliary knowledge. Moreover, as there are no standard datasets and benchmarks to compare visual transfer learning tasks that use KGs, we refer to **RQ4** and provide a list of datasets and benchmarks that have been used in the community of knowledge-based ML and visual transfer learning in Table 3. These Datasets and Benchmarks include: a) Attribute augmented image datasets with textual image or class attribute descriptions; b) Language augmented image datasets, providing additional textual descriptions of the images; c) Knowledge graph augmented image datasets, containing meta information of class relations in a KG; d) Image datasets without auxiliary knowledge, used for zero-shot learning and domain generalization tasks.

5.1. *Generic knowledge graphs*

Over the years, several open-access KGs have been created by various community initiatives. These graphs contain universal knowledge which potentially can be used as auxiliary knowledge in various scenarios. In the following, some of the most common generic KGs currently available are described in more detail. However, for deeper insights, we refer to the survey of Färber et al. [34].

WordNet [92]: WordNet, firstly released in 1995, is an online lexical reference system for English nouns, verbs, and adjectives which are organized into *synonym sets* (synsets), each representing one underlying lexical concept. WordNet superficially resembles a thesaurus, in that it groups words based on their meanings. There are 117,000 synsets, each synset is linked with other synsets by super-subordinate relations, forming a hierarchical structure of instances, concepts and categories whereas all are linked with the root node, *entity*.

ConceptNet 5.5 [129]: ConceptNet 5.5 is a KG that connects words and phrases of natural language with labeled edges. Its knowledge is collected from many sources that include expert-created resources, crowd-sourcing, and games with a purpose. It is designed to represent the general knowledge involved in understanding language, improving natural language applications by allowing the application to better understand the meanings behind the words people use. Information within ConceptNet is modeled as a directed labeled graph (see Section 3.2), where concepts are connected via binary relationships. It contains approximately 34 million statements, i.e. edges.[8]

[8] https://conceptnet.io

Table 3

Datasets and benchmarks of the field of visual transfer learning and knowledge-based ML are summarized due to type of knowledge, task, auxiliary knowledge, and their release date. ZSL is zero-shot-learning, DG is domain generalization, and other are tasks from image classification, object detection, object segmentation, and image captioning

Type of knowledge	Task	Dataset	Auxiliary knowledge	Release date
Attributes + Images	ZSL	AwA	textual attributes for img/cls	2009
		AwA2	textual attributes for img/cls	2019
		SUN	textual attributes for img/cls	2012
		CUB	textual attributes for img/cls	2010
	DG	Large-Scale Car Dataset	textual attributes for img/cls	2017
Language + Images	Other	MS-COCO	textual denotation graph	2014
		Flickr30K	textual denotation graph	2015
		SBU Captions	textual descriptions for img	2011
		Conceptual Captions	textual descriptions for img	2018
Knowledge Graph + Images	ZSL	Visual Genome	flat concept graph	2017
		miniImageNet	hierarchical concept graph	2016
		tiredImageNet	hierarchical concept graph	2018
	DG	ImageNet	hierarchical concept graph	2009–2015
Images	ZSL	CIFAR-FS	N/A	2016
		FC-100	N/A	2016
	DG	Office-31	N/A	2010
		Office-Home	N/A	2016
		VisDA2017	N/A	2017

DBPedia [5]: DBPedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. The underlying structure of DBpedia is a hypergraph model (see Section 3.2) where facts are represented via binary and n-ary relationships. The English version of the DBpedia knowledge base describes 4,58 million things, out of which 4,22 million are classified in a consistent ontology, including 1,445,000 persons, 735,000 places, and 411,000 creative works.[9]

Wikidata [140]: Wikidata is a KG, built collaboratively by humans or automated agents. It encapsulates facts about the world entities organized in a form of complex statements. The basic structure comprises items defined with a label and several aliases. In addition, Wikidata contains some sense of basic commonsense knowledge [60] which allows for performing several sophisticated downstream tasks based on reasoning capabilities. The facts within Wikidata are represented as a hyper-relation graph (see Section 3.2) where relations are enriched with additional information known as qualifiers [40]. These qualifiers enable the disambiguation of complex facts about the same entities in different contexts. Currently, Wikidata has 92,4 million items, where around 6,3 million of them are humans, 2 million administrative entities, 22,5 million scholarly articles, and so on.[10]

5.2. Image datasets with auxiliary knowledge

Some datasets are built on auxiliary knowledge bases or intended to use with auxiliary information. We provide a categorization of the datasets and benchmarks concerning the type of auxiliary knowledge it is augmented with.

5.2.1. Attribute augmented image datasets

Attribute augmented image datasets are image datasets with additional descriptions of image and class attributes, used for knowledge-based ML.

[9] https://wiki.dbpedia.org/about

[10] https://www.wikidata.org/wiki/Wikidata:Statistics, accessed on 02 February 2021.

AwA [76]: The *Animals with Attributes* dataset consists of over 30,000 images with pre-computed reference features for 50 animal classes, for which a semantic attribute annotation is available from studies in cognitive science. However, as the AWA images do not have a public copyright license, only some computed image features, i.e. SIFT [87], DECAF [33], VGG19 [126] of AWA dataset are publicly available, rather than the raw images. Since image feature learning is an important part of modern CV, this dataset is of limited use for end-to-end learned visual models.

AwA2 [153]: The *Animals with Attributes 2* dataset is recently introduced and has roughly the same number of images all with public licenses, and the same number of classes and attributes as the AwA dataset.

CUB [150]: The *Caltech-UCSD-Birds 200-2011* dataset is a fine-grained and medium scale dataset concerning both the number of images and the number of classes, i.e. 11,788 images from 200 different types of birds annotated with 312 attributes. Akata et al. [2] introduces the first zero-shot split of CUB with 150 training, 50 validation, and 50 test classes.

SUN [104]: The *Scene Categorization Benchmark* is also a fine-grained and medium-sized dataset, both in terms of the number of images and the number of classes., i.e. SUN contains 14,340 images coming from 717 types of scenes annotated with 102 attributes. Lampert et al. [77] use 645 classes of SUN for training, 65 classes for validation, and 72 classes for testing.

Large-scale car dataset [43]: The *Large-Scale Car Dataset* originally consists of 2,657 classes and 1,095,021 images from four sources: craigslist.com, cars.com, edmunds.com and Google Street View. They refer to images from craigslist.com, cars.com and edmunds.com as web images and those from Google Street View as GSV images. It was adapted to domain generalization using a subset of 170 classes and 71,030 images [42]. The image category web images is used as source domain, whereas the category GSV images suits as target domain. The cars in web images are large and typically un-occluded whereas those in GSV are small, blurry and occluded. In addition to the category labels, each class is accompanied by metadata such as the make, model body type, and manufacturing country of the car.

5.2.2. *Language augmented image datasets*

These image datasets are enriched with additional textual descriptions and captions of images. To categorize images based on the textual descriptions, denotation graphs are introduced and are available for some datasets.

MS-COCO [83]: *MS-COCO* includes images of complex everyday scenes with common objects in their natural context. It contains a total of 2.5 million labeled instances of 91 object types, in 328k images, each accompanied with five human-written captions. It is used for category detection, instance spotting, and instance segmentation. Recently, Zhang et. al [159] released an additionally learned denotation graph for MS-COCO, which induces a partial ordering over the textual image descriptions. There is also work that extends *MS-COCO* to zero-shot learning tasks by providing additional splits of unseen and seen class categories [8].

Flickr30K [156]: The *Flickr30K* is a standard benchmark for sentence-based image description and was originally developed for the tasks of image-based and text-based retrieval. The dataset contains 31K images collected from the Flickr website, with five textual descriptions per image. Each image is described independently by five annotators who are not familiar with the specific entities and circumstances, resulting in high-level descriptions such as “Three people setting up a tent”. The images are under the Creative Commons license. Moreover, they released a denotation graph for the dataset [159].

SBU captions [100]: *SBU Captions* contains a large number of images from the Flickr website. They are filtered to produce a data collection containing over 1 million well-captioned images. The images have rich user-associated captions from a web-scale captioned image collection. These text descriptions generally work similarly to captions and usually relate directly to some aspect of the visual image content.

Conceptual captions [125]: *Conceptual Captions* consists of an order of magnitude more images than the MS-COCO dataset and represents a wider variety of both images and image caption styles. Therefore, they extracted and filtered image caption annotations from billions of internet sources, e.g. webpages.

5.2.3. *Knowledge graph augmented image datasets*

These datasets are augmented with an additional KG describing relations between classes or a scene in an image.

Visual genome [74]: *Visual Genome* provides a flat concept graph model of object relationships in images. Dense annotations of objects, attributes, and relationships within each image are collected. Specifically, the dataset contains over 100K images where each image has an average of 21 objects, 18 attributes, and 18 pairwise relationships between objects. For zero-shot learning a split with 608 categories are considered for classification [8,88]. Among these, 478 are seen categories, and 130 are unseen categories. This results in 54,913 training images and 7,788 test images. The relationship graph in the dataset has 6,396 edges.

ImageNet [119]: The *ImageNet Large-Scale Visual Recognition Dataset and Challenge* is a benchmark in object category classification and detection on hundreds of categories and millions of images. The challenge has been run annually from 2010 to 2015. It contains 1000 classes and more than 1,2 mil train, and 100K test images per class for object classification. For object detection, it contains 1000 classes and more than 450K training images with 470K bounding boxes, 50K validation images with 55K bounding boxes, and 40K test images per class.

There are several derivatives of ImageNet with different appearances, as *ImageNetV2* [111], *ImageNet Sketch* [142], *ImageNet-Vid* [124], *ImageNet Adversarial* [56], *ImageNet Rendition* [54], and such with synthetic distribution shifts, as *ImageNet-C* [55], and *Stylized ImageNet* [44]. More recently, a domain generalization scenario has been created in which ImageNet-trained models are tested on various ImageNet derivatives to evaluate the robustness of the models to distribution shift.

MiniImageNet [138]: *MiniImageNet* is a derivative of the ImageNet dataset and consisting of 60K color images of size 84 × 84 with 100 classes, each having 600 examples. Since this dataset fits in memory on modern computers, it is very convenient for rapid prototyping and experimentation. These 100 classes are divided into 64 train, 16 val, and 20 test classes for the zero-shot learning task.

TiredImageNet [112]: *TiredImageNet* is a subset of the ImageNet dataset. It groups classes into broader categories corresponding to higher-level nodes in the ImageNet hierarchy. There are 34 categories in total, with each category containing between 10 and 30 classes. For zero-shot learning they split the categories into 20 training, 6 validation, and 8 testing categories. This ensures that all of the training classes are sufficiently distinct from the testing classes, unlike miniImageNet.

5.3. *Image datasets without auxiliary knowledge*

This section introduces transfer learning image datasets that have been originally created without auxiliary knowledge.

5.3.1. *Zero-shot learning datasets without auxiliary knowledge*

We introduce image datasets that have been applied mainly for zero-shot learning or few-shot learning tasks.

CIFAR-FS [10]: *CIFAR-FS* is randomly sampled from CIFAR-100 [75]. CIFAR-100 contains 600 images in each of 100 classes, which are further grouped into 20 superclasses. The limited original resolution of 32 × 32 makes the task harder and at the same time allows fast prototyping. Moreover, the dataset is used for the task of few-shot learning.

FC100 [101]: *Fewshot-CIFAR100* is a derivative of the CIFAR-100 dataset and provides a few-shot learning split of the full CIFAR-100 dataset. The dataset is split into superclasses, rather than into individual classes to minimize the information overlap. Thus the train split contains 60 classes belonging to 12 superclasses, the validation and test contain 20 classes belonging to 5 superclasses each.

5.3.2. *Domain generalization datasets without auxiliary knowledge*

We provide a summary of image datasets that have been applied mainly for domain generalization or domain adaptation tasks.

Office-31 [120]: *Office-31* is an object recognition dataset which contains 31 categories and three domains, that is, *Amazon* (A), *Webcam* (W), and *DSLR* (D). These three domains have 2817, 498, and 795 instances, respectively. The images in Amazon are the online e-commerce images taken from Amazon.com. The images in Webcam are the low-resolution images taken by web cameras. And the images in DSLR are the high-resolution images taken by DSLR cameras. In the experiments, every two of the three domains are selected as the source and the target domains, which results in six tasks. The evaluation contains all 6 cross-domain tasks: A → D, A → W, D → A, D → W, W → A,W → D.

Office-home [137]: *Office Home* contains 15,585 images of 65 categories, collected from 4 domains: a) Art: 2421 artistic depictions of objects in the form of sketches, paintings, ornamentation, etc.; b) Clipart: a collection of 4379 clipart images; c) Product: 4428 images of objects without a background, akin to the Amazon category in Office dataset; d) Real-World: 4357 images of objects captured with a regular camera. The evaluation contains all 12 cross-domain tasks.

VisDA2017 [105]: *The 2017 Visual Domain Adaptation Dataset and Challenge* is focused on the simulation-to-reality shift and has two associated tasks: image classification and image segmentation. The goal in both tracks is to first train a model on simulated, synthetic data in the source domain and then adapt it to perform well on real image data in the unlabeled test domain. VisDA2017 is the largest dataset for cross-domain object classification, with over 280K images across 12 categories in the combined training, validation, and testing domains. The image segmentation dataset is also large-scale with over 30K images across 18 categories in the three domains.

6. Related surveys

Since our survey explores approaches that are at the intersection of visual transfer learning and knowledge-based machine learning, we look at well-known surveys from both fields in this section. Furthermore, we provide additional insight into surveys on the topic of explainable AI, as the field is strongly related to knowledge-based ML.

Visual transfer learning: Pan et al. [103] and Zhang et al. [161] categorized the task of visual transfer learning into three main settings: inductive, transductive, and unsupervised transfer learning. In inductive transfer learning the task changes from source to target, whereas the domain stays the same. In transductive transfer learning, the source and target tasks are the same, while the source and target domains are different. Finally, in the unsupervised transfer learning setting, similar to inductive transfer learning, the target task is different from but related to the source task. However, unsupervised transfer learning focuses on solving learning tasks when no labeled data is available in the source and the target domain. Weiss et al. [149] separated the field into homogeneous and heterogeneous transfer learning, whereas approaches of the former are developed and proposed for handling the situations where the domains are of the same feature space and the latter refers to the knowledge transfer process in the situations where the domains have different feature spaces. Kaboli et al. [66] reviewed and structured 20 transfer-learning approaches. Wang et al. [144] investigated the field from the domain change perspective. If the domain change is small they call it homogeneous transfer learning and if the domain change is large they call it heterogeneous transfer learning. Zhang et al. [160] further separated the field of transfer learning into 17 different tasks, based on supervision, the amount of labeled data, and the size of the domain gap. Zhang et al. [161] categorized transfer learning based on their adaptation process into weakly supervised learning, instance re-weighting, feature adaptation, classifier adaptation, deep network adaptation, and adversarial adaptation. Wang et al. [146] provide a comprehensive survey about zero-shot learning methods and their different semantic spaces. These semantic spaces can either be engineered semantic spaces, generated by attributes, lexicals, and text-keywords, or learned semantic spaces, as label-embeddings, text-embeddings, and image-representations. Xian et al. [153] recently released a survey about zero-shot learning where they structured the field into methods that learn linear compatibility, nonlinear compatibility, intermediate attribute classifier, or hybrid models.

Knowledge-based machine learning: Only a few surveys have investigated the field of knowledge-based ML. Von Rueden et al. [139] recently published a survey about knowledge-based ML under the term *informed machine learning*. They structure the field based on the source of the knowledge, the representation of the knowledge, and the integration of the knowledge into the ML pipeline. Further, Gouidis et al. [50] structured the knowledge-based ML literature into approaches with symbolic knowledge, commonsense knowledge, and the ability to learn new knowledge. They give an overview of different works that combines ML with knowledge-based approaches in the field of CV. They categorized the approaches due to their CV task, e.g. object detection, scene understanding, image classification, their applied ML architecture, e.g. CNN, GNN, RCNN, and their loss function, e.g. scoring functions, probabilistic programming models, Bayesian Networks. Ding et al. [32] reviewed all ontology applications in the field of object recognition. Another research field in demand is *Explainable AI*, where knowledge-based methods and ML approaches are combined. Explainable AI refers to methods and techniques of ML such that the results of the solution can be understood by humans. Futia et al. [39] investigated the field of explainable AI using KGs and categorized approaches into knowledge matching, cross-disciplinary and interactive explanations. Chen et al. [20] and Chari et al. [18] proposed to use hybrid explanations of a taxonomy generated for the end-user, including causal methods, neuro-symbolic AI systems, and representation techniques. Seeliger et al. [122] summarized semantic web technologies that can provide valid explanations for ML models, separating them due to their ML technique and semantic expressiveness. Chen et al. [19] recently proposed a survey about knowledge-aware zero-shot learning. They divided the machine learning methods that approach the zero-shot learning task into three distinct categories: mapping function based, generative model based, and graph neural network based. They provided an overview of different types of auxiliary knowledge, e.g. text, attribute, knowledge graph, and rule and ontology.

Aditya et al. [1] provide a survey about reasoning mechanisms and knowledge integration methods for image understanding applications.

Besides an overview of frameworks that handle logic operations, they briefly discuss at which position auxiliary knowledge can be introduced into a DL pipeline: i) Ahead of the DNN, through a pre-processing of domain knowledge and augmentation of training samples; ii) Inside the DNN, through a vectorization of parts of the knowledge base and as an input to intermediate layers; iii) Inside of the DNN, to inspire the neural network architecture; and iv) After the DNN, as a post-processing using external knowledge. We understand their taxonomy as a general explanation of where external knowledge can be induced into the DL pipeline. For instance, our category *Knowledge Graph as a Reviewer* is related to iv), since the KG can operate as a post-processing network on the output of the visual DNN. However, we also see that the reasoning process of the *Knowledge Graph as a Reviewer* can be applied on an intermediate visual feature layer of the DNN. Similarly, the categories *Knowledge Graph as a Trainee*, *Knowledge Graph as a Trainer*, and *Knowledge Graph as a Peer* have overlaps with categories ii) and iii). However, in contrast to Aditya et al. our categories are described by the explicit information exchange between the visual and semantic embedding space. Instead of a categorization based on the position of the knowledge induction, our categories depend on whether the semantic embedding inspires the visual embedding or vice versa. Using our categories, we therefore describe four distinct principles used to combine the two modalities.

Our survey explores the field of visual transfer learning using KGs. Rather than just structuring the field, we also aim to provide the necessary tools for using KGs with DL pipelines to facilitate a straightforward entry. Therefore, we present different modeling structures for KGs, concepts about visual and semantic feature extractors, and different methods for converting KGs into a vector-based h_s. The main contribution is a categorization into four distinct categories of how a KG can be used with a DL pipeline for visual transfer learning tasks. To enable a fair comparison for approaches of visual transfer learning using KGs, we summarize available KGs, datasets, and benchmarks.

7. Challenges and open issues

Integrating auxiliary knowledge in form of a KG into the DL pipeline not only helps in tackling challenges such as catastrophic forgetting or the need for a huge amount of data in transfer learning scenarios, but it also improves the robustness of DL approaches against naturally occurring domain shift. However, exploiting this type of knowledge brings up new challenges related to knowledge representation and utilization, which we are going to discuss in the following.

Relevant knowledge and its representation: A major challenging task when dealing with modeling the knowledge for a given domain is to analyze what type of knowledge is relevant for performing a given task. Currently, the majority of approaches focus on exploiting only the type of knowledge that is truly irrelevant to the context. Furthermore, the temporal aspects between pieces of knowledge are minimally exploited or not exploited at all. As described in Section 3.2, various modeling structures exist that can be used to represent multidimensional information. However, the difficulty raised here is keeping the trade-off between the relevant knowledge and complexity of structures used to represent that.

Evolving knowledge: In daily scenarios, CV-related applications based on ML consume an abundant amount of data collected from various sensors. Typically, this information is used for training purposes in form of vectors performing complex calculations to learn mathematical functions that best fit downstream tasks. A crucial challenge here is to extract and integrate heterogeneous knowledge that can be managed and refined by humans. Progress in the field of KG construction by embedding methods of language and information extraction has already been achieved. [30,31,70]. This would enable the definition of different complex rules and reusable knowledge structures which later can be incorporated back to the existing or new ML pipelines.

Knowledge embedding methods: As we pointed out in Section 3.3, there is a strong relation between knowledge graph embeddings and language embeddings as both are generated by a semantic feature extractor. Using this assumption, we can apply knowledge graph embeddings in various new domains, where language embeddings have shown great potential, with the advantage that h_s can be manually adapted to our needs. This is done either by refining the knowledge in a KG or by using a particular embedding method relevant to the graph structure to best represent the inherent knowledge. The challenge here is related to find suitable KGs and their modeling techniques to form either task-specific or universal h_s spaces that support and enhance DL approaches in CV.

Joint embedding learning: We have seen that basic supervised learning methods that use CE tend to overfit the training data, leading to extensive problems when applied scenarios with a domain shift. Finding a good embedding space is crucial which would enable it to be applied to multiple downstream tasks. To learn efficiently on high dimensional spaces, energy-based functions instead of maximum likelihood seem to be promising, which should be further investigated under different requirements, like imbalance distribution within datasets. As described in Section 3.5, the quality of the combination of visual and semantic embedding space is highly dependent on the similarity measure, the training objective, and the optimization method. It is still an open challenge how to best fit these three parameters to find accurate combinations for a joint embedding space. Moreover, learning visual features extractors directly on semantic embedding spaces with other features, e.g., temporal or contextualized embeddings, instead of discrete labels is a major challenge for future research.

8. Discussion and conclusion

Visual transfer learning using different types of auxiliary knowledge has gained increasing attention in research. Since initiatives for building and maintaining generic knowledge graphs host a large research community, we believe that exploiting them with DL will improve various applications, especially in visual transfer learning. The insights gained in this survey can be useful to conceive solutions for addressing the identified challenges and open issues.

The survey investigates various forms of how KGs as a unified representation of auxiliary knowledge can be used based on a deep analysis of existing approaches. Different graph models, corresponding embedding methods, and suitable training objectives to operate on high-dimensional spaces are described in detail. The major contributions of the survey are formulated in four research questions presented in Section 2. The answers to these questions are given as follows:

- **RQ1** – *How can a knowledge graph be combined with a deep learning pipeline?*
  Approaches of the field of visual transfer learning using KG can be separated into four distinct categories based on how the KG is combined with the DL pipeline:
  1) *Knowledge Graph as a Reviewer* – where the KG is used for post-validation of a visual model;

2) *Knowledge Graph as a Trainee*, where a semantic-visual embedding $h_{s,v}$ is learned using a visual embedding h_v as objective;
3) *Knowledge Graph as a Trainer*, a visual-semantic embedding $h_{v,s}$ is learned using a semantic embedding h_s as objective; and
4) *Knowledge Graph as a Peer*, where a hybrid-embedding h_h is learned using a combination of semantic embedding h_s and a visual embedding h_v as objective.

- **RQ2** – *What are the properties of the respective combinations?* It can be seen that every category has its applications in distinct tasks.
1) *Knowledge Graph as a Reviewer* – approaches leverage auxiliary knowledge by using it as an independent post-validation. The KG or h_s enables reasoning over the output or intermediate feature layers of the DNN. However, the modalities are either learned independently or in sequential order, so that semantic and visual embedding space are not directly influenced by each other.
2) *Knowledge Graph as a Trainee* – approaches leverage auxiliary knowledge by providing a structure for a KGE-Method, e.g. GNN, that is learned using h_v as objective. Approaches are used mainly in the zero-shot learning scenario to extend the learned model to classes that are not present in the training data, using the inductive property of GNNs combined with the ability of DNNs to extract relevant features of images.
3) *Knowledge Graph as a Trainer* – approaches leverage auxiliary knowledge by influencing DNNs in learning specific visual features. The DNN can learn an image data distribution independent embedding provided by h_s instead of just using the data distribution. Thus, we see the advantage of these approaches specifically in the domain generalization scenario.
4) *Knowledge Graph as a Peer* – approaches leverage auxiliary knowledge by influencing semantic and visual embedding equally. Although it is not clear which modality dominates the other and therefore the learned embedding, approaches have yielded quite promising results for zero-shot learning and domain generalization tasks.
- **RQ3** – *Which knowledge graphs already exist, that can be used as auxiliary knowledge?* We provide a short overview of generic KGs that could be used as a basis to form either specific or general approaches for the task of visual transfer learning using KGs.
WordNet, an online lexical reference system for English nouns, verbs, and adjectives, often used to build hierarchical relationship graphs of classes in the image dataset.
ConceptNet 5.5, a commonsense KG that connects words and phrases of natural language, often used to provide flat relationships between different classes of the image dataset.
DBPedia, a KG that represents structured information from Wikipedia and therefore allows to extract facts.
Wikidata, a commonsense KG built collaboratively by humans or automated agents with reasoning capabilities.
- **RQ4** – *What datasets exist, that can be used in the combination with auxiliary knowledge to evaluate visual transfer learning?* We present several vision datasets and cluster them based on the type of auxiliary data they are augmented with.
Attribute Augmented Image Datasets, as Awa, Awa2, CUB, SUN, and Large-Scale Car Dataset.
Language Augmented Image Datasets, as MS-COCO, Flickr30K, SBU Captions, and Conceptual Captions.
Knowledge Graph Augmented Image Datasets, as Visual Genome, ImageNet, miniImageNet, and tiredImageNet.
Image Datasets without Auxiliary Knowledge for zero-shot learning, as CIFAR-FS, FC100, or domain generalization, as Office-31, Office-Home, and VisDA2017.

Future work is directed on conducting extensive experiments using KGs for visual transfer learning tasks while measuring various metrics, such as precision, recall, and accuracy. Furthermore, it will be relevant to investigate the impact of knowledge structures represented via the three common graph models, the impact of different KGE-Methods, and the impact of the four categories a KG can be combined with the DL pipeline on the metrics as above. We hope that this survey will help the reader to combine the technology of KGs and DL to develop models that can benefit from the appropriate combination of visual information with underlying semantic information.

Acknowledgements

This publication was created as part of the research project “KI Delta Learning” (project number: 19A19013D) funded by the Federal Ministry for Economic Affairs and Energy (BMWi) on the basis of a decision by the German Bundestag.

References

[1] S. Aditya, Y. Yang and C. Baral, Integrating knowledge and reasoning in image understanding, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, Macao, China, August 10–16, 2019, S. Kraus, ed., ijcai.org, 2019, pp. 6252–6259. doi:10.24963/ijcai.2019/873.

[2] Z. Akata, F. Perronnin, Z. Harchaoui and C. Schmid, Label-embedding for image classification, *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(7) (2016), 1425–1438. doi:10.1109/TPAMI.2015.2487986.

[3] R. Angles and C. Gutierrez, An introduction to graph data management, in: *Graph Data Management, Fundamental Issues and Recent Developments*, G.H.L. Fletcher, J. Hidders and J. Larriba-Pey, eds, Data-Centric Systems and Applications, Springer, 2018, pp. 1–32. doi:10.1007/978-3-319-96193-4_1.

[4] R. Angles, H. Thakkar and D. Tomaszuk, Mapping RDF databases to property graph databases, *IEEE Access* **8** (2020), 86091–86110. doi:10.1109/ACCESS.2020.2993117.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z.G. Ives, DBpedia: A nucleus for a web of open data, in: *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007*, Busan, Korea, November 11–15, 2007, K. Aberer, K. Choi, N.F. Noy, D. Allemang, K. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber and P. Cudré-Mauroux, eds, Lecture Notes in Computer Science, Vol. 4825, Springer, 2007, pp. 722–735. doi:10.1007/978-3-540-76298-0_52.

[6] L.J. Ba, K. Swersky, S. Fidler and R. Salakhutdinov, Predicting deep zero-shot convolutional neural networks using textual descriptions, in: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, December 7–13, 2015, IEEE Computer Society, 2015, pp. 4247–4255. doi:10.1109/ICCV.2015.483.

[7] I. Balazevic, C. Allen and T.M. Hospedales, TuckER: Tensor factorization for knowledge graph completion, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, Hong Kong, China, November 3–7, 2019, K. Inui, J. Jiang, V. Ng and X. Wan, eds, Association for Computational Linguistics, 2019, pp. 5184–5193. doi:10.18653/v1/D19-1522.

[8] A. Bansal, K. Sikka, G. Sharma, R. Chellappa and A. Divakaran, Zero-shot object detection, in: *Computer Vision – ECCV 2018 – 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part I*, Munich, Germany, September 8–14, 2018, V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss, eds, Lecture Notes in Computer Science, Vol. 11205, Springer, 2018, pp. 397–414. doi:10.1007/978-3-030-01246-5_24.

[9] I.M. Baytas, C. Xiao, F. Wang, A.K. Jain and J. Zhou, Heterogeneous hyper-network embedding, in: *IEEE International Conference on Data Mining, ICDM 2018*, Singapore, November 17–20, 2018, IEEE Computer Society, 2018, pp. 875–880. doi:10.1109/ICDM.2018.00104.

[10] L. Bertinetto, J.F. Henriques, P.H.S. Torr and A. Vedaldi, Meta-learning with differentiable closed-form solvers, in: *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6–9, 2019, OpenReview.net, 2019. https://openreview.net/forum?id=HyxnZh0ct7.

[11] C.M. Bishop, *Pattern Recognition and Machine Learning*, 5th edn, Information Science and Statistics, Springer, 2007, https://www.worldcat.org/oclc/71008143. ISBN 9780387310732.

[12] G. Blanchard, G. Lee and C. Scott, Generalizing from several related classification tasks to a new unlabeled sample, in: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, Proceedings of a Meeting Held*, Granada, Spain, 12–14 December 2011, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F.C.N. Pereira and K.Q. Weinberger, eds, 2011, pp. 2178–2186, https://proceedings.neurips.cc/paper/2011/hash/b571ecea16a9824023ee1af16897a582-Abstract.html.

[13] M. Boudiaf, J. Rony, I.M. Ziko, E. Granger, M. Pedersoli, P. Piantanida and I.B. Ayed, A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses, in: *Computer Vision – ECCV 2020 – 16th European Conference*, Glasgow, UK, August 23–28, 2020, A. Vedaldi, H. Bischof, T. Brox and J. Frahm, eds, Lecture Notes in Computer Science, Vol. 12351, Springer, 2020, pp. 548–564. doi:10.1007/978-3-030-58539-6_33.

[14] J.S. Bridle, Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition, in: *Neurocomputing – Algorithms, Architectures and Applications, Proceedings of the NATO Advanced Research Workshop on Neurocomputing Algorithms, Architectures and Applications*, Les Arcs, France, February 27–March 3, 1989, F. Fogelman-Soulié and J. Hérault, eds, NATO ASI Series, Vol. 68, Springer, 1989, pp. 227–236. doi:10.1007/978-3-642-76153-9_28.

[15] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré and K. Murphy, Machine Learning on Graphs: A Model and Comprehensive Taxonomy, *CoRR* (2020), abs/2005.03675.

[16] I. Chami, Z. Ying, C. Ré and J. Leskovec, Hyperbolic graph convolutional neural networks, in: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, Vancouver, BC, Canada, December 8–14, 2019, H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox and R. Garnett, eds, 2019, pp. 4869–4880, https://proceedings.neurips.cc/paper/2019/hash/0415740eaa4d9decbc8da001d3fd805f-Abstract.html.

[17] S. Changpinyo, W. Chao, B. Gong and F. Sha, Synthesized classifiers for zero-shot learning, in: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, 2016, pp. 5327–5336. doi:10.1109/CVPR.2016.575.

[18] S. Chari, D.M. Gruen, O. Seneviratne and D.L. McGuinness, Directions for explainable knowledge-enabled systems, in: *Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges*, I. Tiddi, F. Lécué and P. Hitzler, eds, Studies on the Semantic Web, Vol. 47, IOS Press, 2020, pp. 245–261. doi:10.3233/SSW200022.

[19] J. Chen, Y. Geng, Z. Chen, I. Horrocks, J.Z. Pan and H. Chen, Knowledge-aware zero-shot learning: Survey and perspective, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event*, Montreal, Canada, 19–27 August 2021, Z. Zhou, ed., ijcai.org, 2021, pp. 4366–4373. doi:10.24963/ijcai.2021/597.

[20] J. Chen, F. Lecue, J. Pan, I. Horrocks and H. Chen, Knowledge-based transfer learning explanation, in: *KR2018 – Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference*, Tempe, United States, 2018, https://hal.inria.fr/hal-01934907.

[21] R. Chen, T. Chen, X. Hui, H. Wu, G. Li and L. Lin, Knowledge graph transfer network for few-shot recognition, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7–12, 2020, AAAI Press, 2020, pp. 10575–10582, https://aaai.org/ojs/index.php/AAAI/article/view/6630.

[22] T. Chen, S. Kornblith, K. Swersky, M. Norouzi and G.E. Hinton, Big self-supervised models are strong semi-supervised learners, in: *NeurIPS*, 2020.

[23] X. Chen, L. Li, L. Fei-Fei and A. Gupta, Iterative visual reasoning beyond convolutions, in: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 7239–7248, http://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Iterative_Visual_Reasoning_CVPR_2018_paper.html. doi:10.1109/CVPR.2018.00756.

[24] Z. Chen, X. Wei, P. Wang and Y. Guo, Multi-label image recognition with graph convolutional networks, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/IEEE, 2019, pp. 5177–5186, http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Multi-Label_Image_Recognition_With_Graph_Convolutional_Networks_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00532.

[25] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, CA, USA, 20–26 June 2005, IEEE Computer Society, 2005, pp. 886–893. doi:10.1109/CVPR.2005.177.

[26] A. D'Amour, K.A. Heller, D. Moldovan, B. Adlam et al., Underspecification Presents Challenges for Credibility in Modern Machine Learning, *CoRR* (2020), https://arxiv.org/abs/2011.03395.

[27] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven and H. Adam, Large-scale object classification using label relation graphs, in: *Computer Vision – ECCV 2014 – 13th European Conference, Proceedings, Part i*, Zurich, Switzerland, September 6–12, 2014, D.J. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, eds, Lecture Notes in Computer Science, Vol. 8689, Springer, 2014, pp. 48–64. doi:10.1007/978-3-319-10590-1_4.

[28] J. Deng, W. Dong, R. Socher, L. Li, K. Li and F. Li, ImageNet: A large-scale hierarchical image database, in: *CVPR*, 2009.

[29] J. Deng, J. Krause, A.C. Berg and F. Li, Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition, in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 16–21, 2012, IEEE Computer Society, 2012, pp. 3450–3457. doi:10.1109/CVPR.2012.6248086.

[30] D. Dessì, F. Osborne, D.R. Recupero, D. Buscaldi and E. Motta, Generating knowledge graphs by employing natural language processing and machine learning techniques within the scholarly domain, *Future Gener. Comput. Syst.* **116** (2021), 253–264. doi:10.1016/j.future.2020.10.026.

[31] A. Dimou, M.V. Sande, P. Colpaert, R. Verborgh, E. Mannens and R.V. de Walle, RML: A generic language for integrated RDF mappings of heterogeneous data, in: *Proceedings of the Workshop on Linked Data on the Web Co-Located with the 23rd International World Wide Web Conference (WWW 2014)*, Seoul, Korea, April 8, 2014, C. Bizer, T. Heath, S. Auer and T. Berners-Lee, eds, CEUR Workshop Proceedings, Vol. 1184, CEUR-WS.org, 2014, http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[32] Z. Ding, L. Yao, B. Liu and J. Wu, Review of the application of ontology in the field of image object recognition, in: *Proceedings of the 11th International Conference on Computer Modeling and Simulation, ICCMS 2019*, North Rockhampton, QLD, Australia, January 16–19, 2019, ACM, 2019, pp. 142–146. doi:10.1145/3307363.3307387.

[33] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell, DeCAF: A deep convolutional activation feature for generic visual recognition, in: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, Beijing, China, 21–26 June 2014, JMLR Workshop and Conference Proceedings, Vol. 32, JMLR.org, 2014, pp. 647–655, http://proceedings.mlr.press/v32/donahue14.html.

[34] M. Färber, B. Ell, C. Menne and A. Rettinger, A comparative survey of dbpedia, freebase, opencyc, wikidata, and yago, *Semantic Web Journal* **1**(1) (2015), 1–5.

[35] B. Fatemi, P. Taslakian, D. Vázquez and D. Poole, Knowledge hypergraphs: Prediction beyond binary relations, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, ed., ijcai.org, 2020, pp. 2191–2197. doi:10.24963/ijcai.2020/303.

[36] Y. Feng, H. You, Z. Zhang, R. Ji and Y. Gao, Hypergraph neural networks, in: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, 2019, pp. 3558–3565. doi:10.1609/aaai.v33i01.33013558.

[37] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato and T. Mikolov, DeViSE: A deep visual-semantic embedding model, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Proceedings of a Meeting Held*, Lake Tahoe, Nevada, United States, December 5–8, 2013, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 2121–2129.

[38] Y. Fu, T.M. Hospedales, T. Xiang and S. Gong, Transductive multi-view zero-shot learning, *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(11) (2015), 2332–2345. doi:10.1109/TPAMI.2015.2408354.

[39] G. Futia and A. Vetrò, On the integration of knowledge graphs into deep learning models for a more comprehensible AI – three challenges for future research, *Inf.* **11**(2) (2020), 122. doi:10.3390/info11020122.

[40] M. Galkin, P. Trivedi, G. Maheshwari, R. Usbeck and J. Lehmann, Message passing for hyper-relational knowledge graphs, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, November 16–20, 2020, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 7346–7359. doi:10.18653/v1/2020.emnlp-main.596.

[41] J. Gao, T. Zhang and C. Xu, I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs, in: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, Honolulu, Hawaii, USA, January 27–February 1, 2019, AAAI Press, 2019, pp. 8303–8311. doi:10.1609/aaai.v33i01.33018303.

[42] T. Gebru, J. Hoffman and L. Fei-Fei, Fine-grained recognition in the wild: A multi-task domain adaptation approach, in: *IEEE International Conference on Computer Vision, ICCV 2017*, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 1358–1367. doi:10.1109/ICCV.2017.151.

[43] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng and L. Fei-Fei, Fine-grained car detection for visual census estimation, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, February 4–9, 2017, S.P. Singh and S. Markovitch, eds, AAAI Press, 2017, pp. 4502–4508, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14583.

[44] R. Geirhos, J. Jacobsen, C. Michaelis, R.S. Zemel, W. Brendel, M. Bethge and F.A. Wichmann, Shortcut Learning in Deep Neural Networks, *CoRR* (2020), abs/2004.07780.

[45] Y. Geng, J. Chen, Z. Chen, J.Z. Pan, Z. Ye, Z. Yuan, Y. Jia and H. Chen, OntoZSL: Ontology-enhanced zero-shot learning, in: *WWW'21: The Web Conference 2021, Virtual Event*, Ljubljana, Slovenia, April 19–23, 2021, J. Leskovec, M. Grobelnik, M. Najork, J. Tang and L. Zia, eds, ACM/IW3C2, 2021, pp. 3325–3336. doi:10.1145/3442381.3450042.

[46] K. Gong, Y. Gao, X. Liang, X. Shen, M. Wang and L. Lin, Graphonomy: Universal human parsing via graph transfer learning, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/IEEE, 2019, pp. 7450–7459, http://openaccess.thecvf.com/content_CVPR_2019/html/Gong_Graphonomy_Universal_Human_Parsing_via_Graph_Transfer_Learning_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00763.

[47] I.J. Goodfellow, Y. Bengio and A.C. Courville, *Deep Learning, Adaptive Computation and Machine Learning*, MIT Press, 2016, http://www.deeplearningbook.org/. ISBN 978-0-262-03561-3.

[48] I.J. Goodfellow, J. Shlens and C. Szegedy, Explaining and harnessing adversarial examples, in: *3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 7–9, 2015, Y. Bengio and Y. LeCun, eds, Conference Track Proceedings, 2015, http://arxiv.org/abs/1412.6572.

[49] M. Gori, G. Monfardini and F. Scarselli, A new model for learning in graph domains, in: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks*, Vol. 2, 2005, pp. 729–734. doi:10.1109/IJCNN.2005.1555942.

[50] F. Gouidis, A. Vassiliades, T. Patkos, A.A. Argyros, N. Bassiliades and D. Plexousakis, A review on intelligent object perception methods combining knowledge-based reasoning and machine learning, in: *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice, AAAI-MAKE 2020, Volume I*, Palo Alto, CA, USA, March 23–25, 2020, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle and F. van Harmelen, eds, CEUR Workshop Proceedings, CEUR-WS.org, 2020, http://ceur-ws.org/Vol-2600/paper7.pdf.

[51] S. Guan, X. Jin, Y. Wang and X. Cheng, Link prediction on N-ary relational data, in: *The World Wide Web Conference, WWW 2019*, San Francisco, CA, USA, May 13–17, 2019, L. Liu, R.W. White, A. Mantrach, F. Silvestri, J.J. McAuley, R. Baeza-Yates and L. Zia, eds, ACM, 2019, pp. 583–593. doi:10.1145/3308558.3313414.

[52] H. Gui, J. Liu, F. Tao, M. Jiang, B. Norick and J. Han, Large-scale embedding learning in heterogeneous event data, in: *IEEE 16th International Conference on Data Mining, ICDM 2016*, Barcelona, Spain, December 12–15, 2016, F. Bonchi, J. Domingo-Ferrer, R. Baeza-Yates, Z. Zhou and X. Wu, eds, IEEE Computer Society, 2016, pp. 907–912. doi:10.1109/ICDM.2016.0111.

[53] R. Hadsell, S. Chopra and Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, New York, NY, USA, 17–22 June 2006, IEEE Computer Society, 2006, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

[54] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt and J. Gilmer, The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization, *CoRR* (2020) abs/2006.16241.

[55] D. Hendrycks and T.G. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, in: *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6–9, 2019, OpenReview.net, 2019. https://openreview.net/forum?id=HJz6tiCqYm.

[56] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt and D. Song, Natural adversarial examples, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual*, June 19–25, 2021, Computer Vision Foundation/IEEE, 2021, pp. 15262–15271, https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

[57] G.E. Hinton and R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, *science* **313**(5786) (2006), 504–507. doi:10.1126/science.1127647.

[58] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J.E.L. Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.N. Ngomo, S.M. Rashid, A. Rula, L. Schmelzeisen, J.F. Sequeda, S. Staab and A. Zimmermann, Knowledge Graphs, *CoRR* (2020), abs/2003.02320.

[59] J. Huang, C. Chen, F. Ye, J. Wu, Z. Zheng and G. Ling, Hyper2vec: Biased random walk for hyper-network embedding, in: *Database Systems for Advanced Applications – 24th International Conference, DASFAA 2019, Proceedings, Part III*, Chiang Mai, Thailand, April 22–25, 2019, *DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Proceedings*, Chiang Mai, Thailand, April 22–25, 2019, G. Li, J. Yang, J. Gama, J. Natwichai and Y. Tong, eds, Lecture Notes in Computer Science, Vol. 11448, Springer, 2019, pp. 273–277. doi:10.1007/978-3-030-18590-9_27.

[60] F. Ilievski, P.A. Szekely and D. Schwabe, *Commonsense Knowledge in Wikidata*, *CoRR* (2020), abs/2008.08114.

[61] M. Jayathilaka, T. Mu and U. Sattler, Ontology-based n-ball concept embeddings informing few-shot image classification, in: *Machine Learning with Symbolic Methods and Knowledge Graphs Co-Located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2021), Virtual*, September 17, 2021, M. Alam, M. Ali, P. Groth, P. Hitzler, J. Lehmann, H. Paulheim, A. Rettinger, H. Sack, A. Sadeghi and V. Tresp, eds, CEUR Workshop Proceedings, Vol. 2997, CEUR-WS.org, 2021, http://ceur-ws.org/Vol-2997/paper1.pdf.

[62] S. Ji, S. Pan, E. Cambria, P. Marttinen and P.S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition and Applications, *CoRR* (2020), abs/2002.00388.

[63] C. Jiang, H. Xu, X. Liang and L. Lin, Hybrid knowledge routed modules for large-scale object detection, in: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS 2018*, Montréal, Canada, December 3–8, 2018, S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, 2018, pp. 1559–1570.

[64] H. Jiang, R. Wang, S. Shan and X. Chen, Transferable contrastive network for generalized zero-shot learning, in: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, Seoul, Korea (South), October 27–November 2, 2019, IEEE 2019, pp. 9764–9773. doi:10.1109/ICCV.2019.00986.

[65] A. Joulin, L. van der Maaten, A. Jabri and N. Vasilache, Learning visual features from large weakly supervised data, in: *Computer Vision – ECCV 2016 – 14th European Conference, Proceedings, Part VII*, Amsterdam, The Netherlands, October 11–14, 2016, B. Leibe, J. Matas, N. Sebe and M. Welling, eds, Lecture Notes in Computer Science, Vol. 9911, Springer, 2016, pp. 67–84. doi:10.1007/978-3-319-46478-7_5.

[66] M. Kaboli, A Review of Transfer LearningAlgorithms, Research Report, Technische Universität München, 2017.

[67] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang and E.P. Xing, Rethinking knowledge graph propagation for zero-shot learning, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/IEEE, 2019, pp. 11487–11496, http://openaccess.thecvf.com/content_CVPR_2019/html/Kampffmeyer_Rethinking_Knowledge_Graph_Propagation_for_Zero-Shot_Learning_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.01175.

[68] A. Karpathy and F. Li, Deep visual-semantic alignments for generating image descriptions, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, June 7–12, 2015, IEEE Computer Society, 2015, pp. 3128–3137. doi:10.1109/CVPR.2015.7298932.

[69] K. Kavukcuoglu, M. Ranzato, R. Fergus and Y. LeCun, Learning invariant features through topographic filter maps, in: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, Florida, USA, 20–25 June 2009, IEEE Computer Society, 2009, pp. 1605–1612. doi:10.1109/CVPR.2009.5206545.

[70] N. Kertkeidkachorn and R. Ichise, An automatic knowledge graph creation framework from natural language text, *IEICE Trans. Inf. Syst.* **101-D(1)** (2018), 90–98. doi:10.1587/transinf.2017SWP0006.

[71] T.N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, in: *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, April 24–26, 2017, Conference Track Proceedings, OpenReview.net, 2017. https://openreview.net/forum?id=SJU4ayYgl.

[72] E. Kodirov, T. Xiang and S. Gong, Semantic autoencoder for zero-shot learning, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 4447–4456. doi:10.1109/CVPR.2017.473.

[73] S. Kornblith, H. Lee, T. Chen and M. Norouzi, What's in a Loss Function for Image Classification? *CoRR* (2020), abs/2010.16402.

[74] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D.A. Shamma, M.S. Bernstein and L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vis.* **123**(1) (2017), 32–73. doi:10.1007/s11263-016-0981-7.

[75] A. Krizhevsky, G. Hinton et al., *Learning Multiple Layers of Features from Tiny Images*, 2009.

[76] C.H. Lampert, H. Nickisch and S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, Miami, Florida, USA, 20–25 June 2009, IEEE Computer Society, 2009, pp. 951–958. doi:10.1109/CVPR.2009.5206594.

[77] C.H. Lampert, H. Nickisch and S. Harmeling, Attribute-based classification for zero-shot visual object categorization, *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(3) (2014), 453–465. doi:10.1109/TPAMI.2013.140.

[78] H. Larochelle, D. Erhan and Y. Bengio, Zero-data learning of new tasks, in: *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008*, Chicago, Illinois, USA, July 13–17, 2008, D. Fox and C.P. Gomes, eds, AAAI Press, 2008, pp. 646–651, http://www.aaai.org/Library/AAAI/2008/aaai08-103.php.

[79] C. Lee, W. Fang, C. Yeh and Y.F. Wang, Multi-label zero-shot learning with structured knowledge graphs, in: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 1576–1585, http://openaccess.thecvf.com/content_cvpr_2018/html/Lee_Multi-Label_Zero-Shot_Learning_CVPR_2018_paper.html. doi:10.1109/CVPR.2018.00170.

[80] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu and Z. Huang, Leveraging the invariant side of generative zero-shot learning, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/IEEE, 2019, pp. 7402–7411, http://openaccess.thecvf.com/content_CVPR_2019/html/Li_Leveraging_the_Invariant_Side_of_Generative_Zero-Shot_Learning_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00758.

[81] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi and J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, in: *Computer Vision – ECCV 2020 – 16th European Conference, Proceedings, Part XXX*, Glasgow, UK, August 23–28, 2020, A. Vedaldi, H. Bischof, T. Brox and J. Frahm, eds, Lecture Notes in Computer Science, Vol. 12375, Springer, 2020, pp. 121–137. doi:10.1007/978-3-030-58577-8_8.

[82] X. Liang, Z. Hu, H. Zhang, L. Lin and E.P. Xing, Symbolic graph reasoning meets convolutions, in: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, Montréal, Canada, December 3–8, 2018, S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, 2018, pp. 1858–1868.

[83] T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C.L. Zitnick, Microsoft COCO: Common objects in context, in: *Computer Vision – ECCV 2014 – 13th European Conference, Proceedings, Part V*, Zurich, Switzerland, September 6–12, 2014, D.J. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars, eds, Lecture Notes in Computer Science, Vol. 8693, Springer, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[84] Y. Liu, Q. Yao and Y. Li, Generalizing tensor decomposition for N-ary relational knowledge bases, in: *WWW'20: The Web Conference 2020*, Taipei, Taiwan, April 20–24, 2020, Y. Huang, I. King, T. Liu and M. van Steen, eds, ACM/IW3C2, 2020, pp. 1104–1114. doi:10.1145/3366423.3380188.

[85] Z. Liu, Z. Jiang and F. Wei, OD-GCN object detection by knowledge graph with GCN, *CoRR* (2019), abs/1908.04385.

[86] V.P.A. Lonij, A. Rawat and M. Nicolae, Open-World Visual Recognition Using Knowledge Graphs, *CoRR* (2017), abs/1708.08310.

[87] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vis.* **60**(2) (2004), 91–110. doi:10.1023/B:VISI.0000029664.99615.94.

[88] R. Luo, N. Zhang, B. Han and L. Yang, Context-aware zero-shot recognition, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7–12, 2020, AAAI Press, 2020, pp. 11709–11716, https://aaai.org/ojs/index.php/AAAI/article/view/6841.

[89] L.V.D. Maaten and G.E. Hinton, Visualizing data using t-SNE, *Journal of Machine Learning Research* **9** (2008), 2579–2605.

[90] K. Marino, R. Salakhutdinov and A. Gupta, The more you know: Using knowledge graphs for image classification, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 20–28. doi:10.1109/CVPR.2017.10.

[91] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado and J. Dean, Distributed representations of words and phrases and their compositionality, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Proceedings of a Meeting Held*, Lake Tahoe, Nevada, United States, December 5–8, 2013, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 3111–3119.

[92] G.A. Miller, WordNet: A Lexical Database for English, *Commun. ACM* (1995).

[93] T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.-M. Chang, V.L. Malave, R.A. Mason and M.A. Just, Predicting human brain activity associated with the meanings of nouns, *Science* **320**(5880) (2008), 1191–1195, https://science.sciencemag.org/content/320/5880/1191. doi:10.1126/science.1152876.

[94] S. Monka, L. Halilaj, S. Schmid and A. Rettinger, Learning visual models using a knowledge graph as a trainer, in: *The Semantic Web – ISWC 2021–20th International Semantic Web Conference, ISWC 2021, Virtual Event, Proceedings*, October 24–28, 2021, A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P.M. Barnaghi, A. Haller, M. Dragoni and H. Alani, eds, Lecture Notes in Computer Science, Vol. 12922, Springer, 2021, pp. 357–373. doi:10.1007/978-3-030-88361-4_21.

[95] K. Muandet, D. Balduzzi and B. Schölkopf, Domain generalization via invariant feature representation, in: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013*, Atlanta, GA, USA, 16–21 June 2013, JMLR Workshop and Conference Proceedings, Vol. 28, JMLR.org, 2013, pp. 10–18, http://proceedings.mlr.press/v28/muandet13.html.

[96] M.F. Naeem, Y. Xian, F. Tombari and Z. Akata, Learning Graph Embeddings for Compositional Zero-shot Learning, *CoRR* (2021), abs/2102.01987.

[97] M.F. Naeem, Y. Xian, F. Tombari and Z. Akata, Learning graph embeddings for compositional zero-shot learning, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual*, June 19–25, 2021, Computer Vision Foundation/IEEE, 2021, pp. 953–962, https://openaccess.thecvf.com/content/CVPR2021/html/Naeem_Learning_Graph_Embeddings_for_Compositional_Zero-Shot_Learning_CVPR_2021_paper.html.

[98] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado and J. Dean, Zero-shot learning by convex combination of semantic embeddings, in: *2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*, Banff, AB, Canada, April 14–16, 2014, Y. Bengio and Y. LeCun, eds, 2014, http://arxiv.org/abs/1312.5650.

[99] V. Ordonez, J. Deng, Y. Choi, A.C. Berg and T.L. Berg, From large scale image categorization to entry-level categories, in: *IEEE International Conference on Computer Vision, ICCV 2013*, Sydney, Australia, December 1–8, 2013, IEEE Computer Society, 2013, pp. 2768–2775. doi:10.1109/ICCV.2013.344.

[100] V. Ordonez, G. Kulkarni and T.L. Berg, Im2Text: Describing images using 1 million captioned photographs, in: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a Meeting Held*, Granada, Spain, 12–14 December 2011, J. Shawe-Taylor, R.S. Zemel, P.L. Bartlett, F.C.N. Pereira and K.Q. Weinberger, eds, 2011, pp. 1143–1151, https://proceedings.neurips.cc/paper/2011/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

[101] B.N. Oreshkin, P.R. López and A. Lacoste, TADAM: Task dependent adaptive metric for improved few-shot learning, in: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, Montréal, Canada, December 3–8, 2018, S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, 2018, pp. 719–729.

[102] M. Palatucci, D. Pomerleau, G.E. Hinton and T.M. Mitchell, Zero-shot learning with semantic output codes, in: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009, Proceedings of a Meeting Held*, Vancouver, British Columbia, Canada, 7–10 December 2009, Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams and A. Culotta, eds, Curran Associates, Inc., 2009, pp. 1410–1418.

[103] S.J. Pan and Q. Yang, A survey on transfer learning, *IEEE Trans. Knowl. Data Eng.* **22**(10) (2010), 1345–1359. doi:10.1109/TKDE.2009.191.

[104] G. Patterson and J. Hays, SUN attribute database: Discovering, annotating, and recognizing scene attributes, in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, June 16–21, 2012, IEEE Computer Society, Providence, RI, USA, 2012, pp. 2751–2758. doi:10.1109/CVPR.2012.6247998.

[105] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang and K. Saenko, VisDA: The Visual Domain Adaptation Challenge, *CoRR* (2017), abs/1710.06924.

[106] Z. Peng, Z. Li, J. Zhang, Y. Li, G. Qi and J. Tang, Few-shot image recognition with knowledge transfer, in: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, Seoul, Korea (South), October 27–November 2, 2019, IEEE, 2019, pp. 441–449. doi:10.1109/ICCV.2019.00053.

[107] J. Pennington, R. Socher and C.D. Manning, Glove: Global vectors for word representation, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, a Meeting of SIGDAT, a Special Interest Group of the ACL*, Doha, Qatar, October 25–29, 2014, A. Moschitti, B. Pang and W. Daelemans, eds, ACL, 2014, pp. 1532–1543. doi:10.3115/v1/d14-1162.

[108] K. Petersen, R. Feldt, S. Mujtaba and M. Mattsson, Systematic mapping studies in software engineering, in: *12th International Conference on Evaluation and Assessment in Software Engineering, EASE 2008*, University of Bari, Italy, 26–27 June 2008, G. Visaggio, M.T. Baldassarre, S.G. Linkman and M. Turner, eds, Workshops in Computing, BCS, 2008, http://ewic.bcs.org/content/ConWebDoc/19543.

[109] K. Petersen, S. Vakkalanka and L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, *Inf. Softw. Technol.* **64** (2015), 1–18. doi:10.1016/j.infsof.2015.03.007.

[110] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., Learning transferable visual models from natural language supervision, *Image* **2** (2021), T2.

[111] B. Recht, R. Roelofs, L. Schmidt and V. Shankar, Do ImageNet classifiers generalize to ImageNet?, in: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, Long Beach, California, USA, 9–15 June 2019, K. Chaudhuri and R. Salakhutdinov, eds, Proceedings of Machine Learning Research, Vol. 97, PMLR, 2019, pp. 5389–5400, http://proceedings.mlr.press/v97/recht19a.html.

[112] M. Ren, E. Triantafillou, S. Ravi, J. Snell, K. Swersky, J.B. Tenenbaum, H. Larochelle and R.S. Zemel, Meta-learning for semi-supervised few-shot classification, in: *6th International Conference on Learning Representations, ICLR 2018, Conference Track Proceedings*, Vancouver, BC, Canada, 30–May 3, 2018, OpenReview.net, 2018, https://openreview.net/forum?id=HJcSzz-CZ.

[113] S. Ren, K. He, R.B. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6) (2017), 1137–1149. doi:10.1109/TPAMI.2016.2577031.

[114] M. Rochan and Y. Wang, Weakly supervised localization of novel objects using appearance transfer, in: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, Boston, MA, USA, June 7–12, 2015, IEEE Computer Society, 2015, pp. 4315–4324. doi:10.1109/CVPR.2015.7299060.

[115] M. Rohrbach, S. Ebert and B. Schiele, Transfer learning in a transductive setting, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Proceedings of a Meeting Held*, Lake Tahoe, Nevada, United States, December 5–8, 2013, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 46–54.

[116] P. Rosso, D. Yang and P. Cudré-Mauroux, Beyond triplets: Hyper-relational knowledge graph embedding for link prediction, in: *WWW'20: The Web Conference 2020*, Taipei, Taiwan, April 20–24, 2020, Y. Huang, I. King, T. Liu and M. van Steen, eds, ACM/IW3C2, 2020, pp. 1885–1896. doi:10.1145/3366423.3380257.

[117] A. Roy, D. Ghosal, E. Cambria, N. Majumder, R. Mihalcea and S. Poria, Improving Zero Shot Learning Baselines with Commonsense Knowledge, *CoRR* (2020), abs/2012.06236.

[118] S. Ruder and B. Plank, Learning to select data for transfer learning with Bayesian Optimization, in: *EMNLP*, 2017.

[119] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M.S. Bernstein, A.C. Berg and F. Li, ImageNet large scale visual recognition challenge, *Int. J. Comput. Vis.* **115**(3) (2015), 211–252. doi:10.1007/s11263-015-0816-y.

[120] K. Saenko, B. Kulis, M. Fritz and T. Darrell, Adapting visual category models to new domains, in: *Computer Vision – ECCV 2010, 11th European Conference on Computer Vision, Proceedings, Part IV*, Heraklion, Crete, Greece, September 5–11, 2010, K. Daniilidis, P. Maragos and N. Paragios, eds, Lecture Notes in Computer Science, Vol. 6314, Springer, 2010, pp. 213–226. doi:10.1007/978-3-642-15561-1_16.

[121] R. Salakhutdinov, A. Torralba and J.B. Tenenbaum, Learning to share visual appearance for multiclass object detection, in: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, Colorado Springs, CO, USA, 20–25 June 2011, IEEE Computer Society, 2011, pp. 1481–1488. doi:10.1109/CVPR.2011.5995720.

[122] A. Seeliger, M. Pfaff and H. Krcmar, Semantic web technologies for explainable machine learning models: A literature review, in: *Joint Proceedings of the 6th International Workshop on Dataset PROFlLing and Search & the 1st Workshop on Semantic Explainability Co-Located with the 18th International Semantic Web Conference (ISWC 2019)*, Auckland, New Zealand, October 27, 2019, E. Demidova, S. Dietze, J.G. Breslin, S. Gottschalk, P. Cimiano, B. Ell, A. Lawrynowicz, L. Moss and A.N. Ngomo, eds, CEUR Workshop Proceedings, Vol. 2465, CEUR-WS.org, 2019, pp. 30–45, http://ceur-ws.org/Vol-2465/semex_paper1.pdf.

[123] B. Shahbaba and R. Neal, Improving classification when a class hierarchy is available using a hierarchy-based prior, *Bayesian Analysis* **2** (2005), 221–237.

[124] V. Shankar, A. Dave, R. Roelofs, D. Ramanan, B. Recht and L. Schmidt, Do image classifiers generalize across time?, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9661–9669.

[125] P. Sharma, N. Ding, S. Goodman and R. Soricut, Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, Melbourne, Australia, July 15–20, 2018, I. Gurevych and Y. Miyao, eds, Association for Computational Linguistics, 2018, pp. 2556–2565, https://www.aclweb.org/anthology/P18-1238/. doi:10.18653/v1/P18-1238.

[126] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, eds, San Diego, CA, USA, May 7–9, 2015, 2015, http://arxiv.org/abs/1409.1556.

[127] R. Socher, D. Chen, C.D. Manning and A.Y. Ng, Reasoning with neural tensor networks for knowledge base completion, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Proceedings of a Meeting Held*, Lake Tahoe, Nevada, United States, December 5–8, 2013, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 926–934, https://proceedings.neurips.cc/paper/2013/hash/b337e84de8752b27eda3a12363109e80-Abstract.html.

[128] R. Socher, M. Ganjoo, C.D. Manning and A.Y. Ng, Zero-shot learning through cross-modal transfer, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Proceedings of a Meeting Held*, Lake Tahoe, Nevada, United States, December 5–8, 2013, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 935–943.

[129] R. Speer, J. Chin and C. Havasi, ConceptNet 5.5: An open multilingual graph of general knowledge, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, San Francisco, California, USA, February 4–9, 2017, S.P. Singh and S. Markovitch, eds, AAAI Press, 2017, pp. 4444–4451, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

[130] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H.S. Torr and T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018, Computer Vision Foundation/IEEE Computer Society, 2018, pp. 1199–1208, http://openaccess.thecvf.com/content_cvpr_2018/html/Sung_Learning_to_Compare_CVPR_2018_paper.html. doi:10.1109/CVPR.2018.00131.

[131] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang and C. Liu, A survey on deep transfer learning, in: *Artificial Neural Networks and Machine Learning – ICANN 2018–27th International Conference on Artificial Neural Networks, Proceedings, Part III*, Rhodes, Greece, October 4–7, 2018, V. Kurková, Y. Manolopoulos, B. Hammer, L.S. Iliadis and I. Maglogiannis, eds, Lecture Notes in Computer Science, Vol. 11141, Springer, Rhodes, Greece, 2018, pp. 270–279. doi:10.1007/978-3-030-01424-7_27.

[132] Y. Tang, J. Wang, X. Wang, B. Gao, E. Dellandréa, R.J. Gaizauskas and L. Chen, Visual and semantic knowledge transfer for large scale semi-supervised object detection, *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(12) (2018), 3045–3058. doi:10.1109/TPAMI.2017.2771779.

[133] Y.H. Tsai, L. Huang and R. Salakhutdinov, Learning robust visual-semantic embeddings, in: *IEEE International Conference on Computer Vision, ICCV 2017*, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 3591–3600. doi:10.1109/ICCV.2017.386.

[134] K. Tu, P. Cui, X. Wang, F. Wang and W. Zhu, Structural deep embedding for hyper-networks, in: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18)*, New Orleans, Louisiana, USA, February 2–7, 2018, S.A. McIlraith and K.Q. Weinberger, eds, AAAI Press, 2018, pp. 426–433, https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16797.

[135] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, Long Beach, CA, USA, December 4–9, 2017, I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan and R. Garnett, eds, 2017, pp. 5998–6008, https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[136] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò and Y. Bengio, Graph attention networks, in: *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 30–May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018, https://openreview.net/forum?id=rJXMpikCZ.

[137] H. Venkateswara, J. Eusebio, S. Chakraborty and S. Panchanathan, Deep hashing network for unsupervised domain adaptation, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 5385–5394. doi:10.1109/CVPR.2017.572.

[138] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu and D. Wierstra, Matching networks for one shot learning, in: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, Barcelona, Spain, December 5–10, 2016, D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon and R. Garnett, eds, 2016, pp. 3630–3638.

[139] L. von Rüden, S. Mayer, J. Garcke, C. Bauckhage and J. Schücker, Informed Machine Learning – Towards a Taxonomy of Explicit Integration of Knowledge into Machine Learning, *CoRR* (2019), abs/1903.12394.

[140] D. Vrandecic, Wikidata: A new platform for collaborative data collection, in: *Proceedings of the 21st World Wide Web Conference, WWW 2012 (Companion Volume)*, Lyon, France, April 16–20, 2012, ACM, 2012, pp. 1063–1064. doi:10.1145/2187980.2188242.

[141] M.R. Vyas, H. Venkateswara and S. Panchanathan, Leveraging seen and unseen semantic relationships for generative zero-shot learning, in: *Computer Vision – ECCV 2020 – 16th European Conference, Proceedings, Part XXX*, Glasgow, UK, August 23–28, 2020, A. Vedaldi, H. Bischof, T. Brox and J. Frahm, eds, Lecture Notes in Computer Science, Vol. 12375, Springer, 2020, pp. 70–86. doi:10.1007/978-3-030-58577-8_5.

[142] H. Wang, P. Lu, H. Zhang, M. Yang, X. Bai, Y. Xu, M. He, Y. Wang and W. Liu, All you need is boundary: Toward arbitrary-shaped text spotting, in: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020*, New York, NY, USA, February 7–12, 2020, AAAI Press, 2020, pp. 12160–12167, https://aaai.org/ojs/index.php/AAAI/article/view/6896.

[143] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen and Y. Wu, Learning fine-grained image similarity with deep ranking, in: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, Columbus, OH, USA, June 23–28, 2014, IEEE Computer Society, 2014, pp. 1386–1393. doi:10.1109/CVPR.2014.180.

[144] M. Wang and W. Deng, Deep visual domain adaptation: A survey, *Neurocomputing* **312** (2018), 135–153. doi:10.1016/j.neucom.2018.05.083.

[145] Q. Wang, Z. Mao, B. Wang and L. Guo, Knowledge graph embedding: A survey of approaches and applications, *IEEE Trans. Knowl. Data Eng.* **29**(12) (2017), 2724–2743. doi:10.1109/TKDE.2017.2754499.

[146] W. Wang, V.W. Zheng, H. Yu and C. Miao, A survey of zero-shot learning: Settings, methods, and applications, *ACM Trans. Intell. Syst. Technol.* **10**(2) (2019), 13:1–13:37. doi:10.1145/3293318.

[147] X. Wang and K.K. Paliwal, Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition, *Pattern Recognit.* **36**(10) (2003), 2429–2439. doi:10.1016/S0031-3203(03)00044-X.

[148] X. Wang, Y. Ye and A. Gupta, Zero-shot recognition via semantic embeddings and knowledge graphs, in: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society, 2018, pp. 6857–6866, http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Zero-Shot_Recognition_via_CVPR_2018_paper.html. doi:10.1109/CVPR.2018.00717.

[149] K.R. Weiss, T.M. Khoshgoftaar and D. Wang, A survey of transfer learning, *J. Big Data* **3** (2016), 9. doi:10.1186/s40537-016-0043-6.

[150] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie and P. Perona, Caltech-UCSD Birds 200, Technical Report, CNS-TR-2010-001, California Institute of Technology, 2010.

[151] J. Wen, J. Li, Y. Mao, S. Chen and R. Zhang, On the representation and embedding of knowledge bases beyond binary relations, in: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*, New York, NY, USA, 9–15 July 2016, S. Kambhampati, ed., IJCAI/AAAI Press, 2016, pp. 1300–1307, http://www.ijcai.org/Abstract/16/188.

[152] C. Wohlin, Guidelines for snowballing in systematic literature studies and a replication in software engineering, in: *18th International Conference on Evaluation and Assessment in Software Engineering, EASE'14*, London, England, United Kingdom, May 13–14, 2014, M.J. Shepperd, T. Hall and I. Myrtveit, eds, ACM, 2014, pp. 38:1–38:10. doi:10.1145/2601248.2601268.

[153] Y. Xian, C.H. Lampert, B. Schiele and Z. Akata, Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly, *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(9) (2019), 2251–2265. doi:10.1109/TPAMI.2018.2857768.

[154] N. Yadati, M. Nimishakavi, P. Yadav, V. Nitin, A. Louis and P.P. Talukdar, in: *HyperGCN: A New Method for Training Graph Convolutional Networks on Hypergraphs, in: Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, 2019, NeurIPS 2019*, Vancouver, BC, Canada, December 8–14, 2019, H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox and R. Garnett, eds, 2019, pp. 1509–1520, https://proceedings.neurips.cc/paper/2019/hash/1efa39bcaec6f3900149160693694536-Abstract.html.

[155] Y. Yang and T.M. Hospedales, A unified perspective on multi-domain and multi-task learning, in: *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, San Diego, CA, USA, May 7–9, 2015, Y. Bengio and Y. LeCun, eds, 2015, http://arxiv.org/abs/1412.7489.

[156] P. Young, A. Lai, M. Hodosh and J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, *Trans. Assoc. Comput. Linguistics* **2** (2014), 67–78, https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229. doi:10.1162/tacl_a_00166.

[157] C. Yu, C. Tai, T. Chan and Y. Yang, Modeling multi-way relations with hypergraph embedding, in: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018*, Torino, Italy, October 22–26, 2018, A. Cuzzocrea, J. Allan, N.W. Paton, D. Srivastava, R. Agrawal, A.Z. Broder, M.J. Zaki, K.S. Candan, A. Labrinidis, A. Schuster and H. Wang, eds, ACM, 2018, pp. 1707–1710. doi:10.1145/3269206.3269274.

[158] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu and H. Wang, ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph, *CoRR* (2020), abs/2006.16934.

[159] B. Zhang, H. Hu, V. Jain, E. Ie and F. Sha, Learning to represent image and text with denotation graph, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online*, November 16–20, 2020, B. Webber, T. Cohn, Y. He and Y. Liu, eds, Association for Computational Linguistics, 2020, pp. 823–839. doi:10.18653/v1/2020.emnlp-main.60.

[160] J. Zhang, W. Li, P. Ogunbona and D. Xu, Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective, *ACM Comput. Surv.* **52**(1) (2019), 7:1–7:38. doi:10.1145/3291124.

[161] L. Zhang, Transfer Adaptation Learning: A Decade Survey, *CoRR* (2019), abs/1903.04687.

[162] L. Zhang, T. Xiang and S. Gong, Learning a deep embedding model for zero-shot learning, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society, 2017, pp. 3010–3019. doi:10.1109/CVPR.2017.321.

[163] R. Zhang, J. Li, J. Mei and Y. Mao, Scalable instance reconstruction in knowledge bases via relatedness affiliated embedding, in: *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*, Lyon, France, April 23–27, 2018, P. Champin, F.L. Gandon, M. Lalmas and P.G. Ipeirotis, eds, ACM, 2018, pp. 1185–1194. doi:10.1145/3178876.3186017.

[164] R. Zhang, Y. Zou and J. Ma, Hyper-SAGNN: A self-attention based graph neural network for hypergraphs, in: *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020. https://openreview.net/forum?id=ryeHuJBtPH.

[165] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning and C.P. Langlotz, Contrastive Learning of Medical Visual Representations from Paired Images and Text, *CoRR* (2020), abs/2010.00747.

[166] Z. Zhang and V. Saligrama, Zero-shot learning via semantic similarity embedding, in: *2015 IEEE International Conference on Computer Vision, ICCV 2015*, Santiago, Chile, December 7–13, 2015, IEEE Computer Society, 2015, pp. 4166–4174. doi:10.1109/ICCV.2015.474.

[167] H. Zhao, X. Puig, B. Zhou, S. Fidler and A. Torralba, Open vocabulary scene parsing, in: *IEEE International Conference on Computer Vision, ICCV 2017*, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 2021–2029. doi:10.1109/ICCV.2017.221.

[168] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso and A. Torralba, Semantic understanding of scenes through the ADE20K dataset, *Int. J. Comput. Vis.* **127**(3) (2019), 302–321. doi:10.1007/s11263-018-1140-0.

[169] Y. Zhu, M. Elhoseiny, B. Liu, X. Peng and A. Elgammal, A generative adversarial approach for zero-shot learning from noisy texts, in: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Salt Lake City, UT, USA, June 18–22, 2018, Computer Vision Foundation/IEEE Computer Society, 2018, pp. 1004–1013, http://openaccess.thecvf.com/content_cvpr_2018/html/Zhu_A_Generative_Adversarial_CVPR_2018_paper.html. doi:10.1109/CVPR.2018.00111.