A Survey on Visual Transfer Learning using Knowledge Graphs

Sebastian Monka [a,b] and 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* (KGE). 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, hyper-relational graphs, and hypergraphs. We explain the notion of feature extractor, while specifically referring to visual and semantic features. We provide a broad overview of KGE-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. 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 [1]. 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 [2]. This has been demonstrated by using adversarially constructed examples [3] or variations in the test images such as noise, blur, and JPEG compression [4]. Moreover, authors in [5] even claim that any standard DNN suffers from such an unpredictable distribution shift when it is deployed in the real world.

1570-0844/21/$35.00 © 2021 – IOS Press and the authors. All rights reserved

Transfer learning is the area of machine learning that tries to find approaches that can deal with such an unpredictable distribution shift [5]. 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 augment 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. [6] auxiliary knowledge is not only important to solve transfer learning problems, but also an opportunity to influence the way a ML model learns from unstructured data.

In this survey, we focus on visual transfer learning approaches using KG, 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* (KGE). The ability to transform the graph-based knowledge into the vector space enables the application of linear operations to KGEs and thus its use in combination with DNNs. Following the successful application of language embeddings, new opportunities are opening up for the use of KGs for CV tasks.

We pursue the goal of enabling the reader to solve visual transfer learning problems with the help of specific KG-DL configurations. Therefore, we first outline different types of modeling structures of knowledge such as directed labeled graphs, hyper-relational graphs, and hypergraphs. Next, we explain the notion of feature extractor, specifically referring to visual and semantic features, where the former mainly includes DL-based visual models and the latter includes language models and KGE-Methods. We provide a broad overview of KGE-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 - the KG is used for post validation of a visual model; 2) Knowledge Graph as a Trainee - the KGE is influenced by the visual embedding; 3) Knowledge Graph as a Trainer - the KGE influences the visual embedding; and 4) Knowledge Graph as a Peer - the KGE and the visual embedding learn a joint embedding space. Due to the shortage of visual transfer learning approaches for category 3) or 4) and their similarities to KGEs, we also considered approaches that use other semantic embedding spaces such as language embeddings as auxiliary knowledge. Furthermore, we provide an overview of generic KGs and several datasets and benchmarks using various types of auxiliary knowledge, like attributes, textual descriptions, or graphs. Last, we summarize related surveys in the field of visual transfer learning and knowledge-based ML and give an outlook about challenges and open issues for future research.

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.

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 [7, 8] and conducted a initial search on five scholarly indexing services. We applied inclusion and exclusion criteria on our findings and extended them based on the snowballing approach [9].

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, there are major challenges and questions that 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** will be 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** will be 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].

2.3. Literature Selection and Quality Assessment

After 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:

- Books and 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 [9], which led to 22 additional studies. On the set of 128 primary studies we conducted 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 general term knowledge in the context of this survey, describes the fundamentals of KGs, feature extractors, knowledge grap embeddings, and joint training objectives.

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 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 [1].

3.1. Knowledge Graph

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 descrip-

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

tions. A comprehensive definition is given by the authors of [10] 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[11]. Each graph model has its own 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. 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.

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 [12]. 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.

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

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.

3.2. 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. 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.

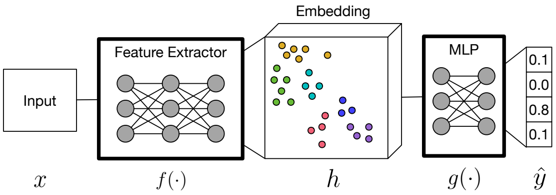


Fig. 1. A DNN that takes $\mathbf{x}$ as input and predicts $\hat{\mathbf{y}}$ can be decoupled into a feature extractor $f(\cdot)$ with its embedding space $\mathbf{h}$ and a prediction task $g(\cdot)$.

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

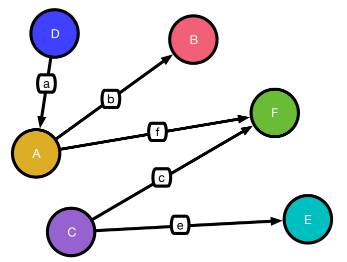
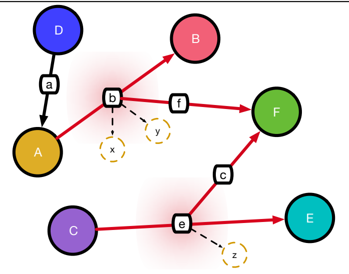
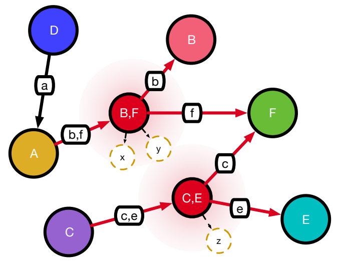
$$\hat{\mathbf{y}} = g(f(\mathbf{x})), \text{ with } f(x) = \mathbf{h}. \tag{1}$$

There are different architectures of DNNs, but they always consist of the same components: neurons, synapses, weights, biases, and functions [1]. The most common architectures that build a DNN are *multi-layer perceptrons* (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 [1].

Whereas, DNNs are usually trained end-to-end resulting in a task-dependent embedding space $\mathbf{h}$, more recently, attempts have been made to independently

Table 1

Various Graph Models. Three common graph models used as underlying structure for knowledge representation in KGs: 1) Directed Labeled Graphs; 2) Hyper-relational Graphs; and 3) Hypergraphs.

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

pre-train the feature extractor that it can be applied to several visual transfer learning and downstream tasks [13].

3.2.1. *Visual Features Extractor*

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

A formal definition is given by

$$\mathbf{h}_v = f_v(\mathbf{x_v}), \tag{2}$$

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

Whereas early approaches used traditional visual features extractors as *scale-invariant feature transform* (SIFT)[14] or *histogram of oriented gradients* (HOG) [15], modern CV methods use almost only DNN-based approaches.

3.2.2. *Semantic Features Extractor*

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

A formal definition is given by

$$\mathbf{h}_s = f_s(\mathbf{x_s}), \tag{3}$$

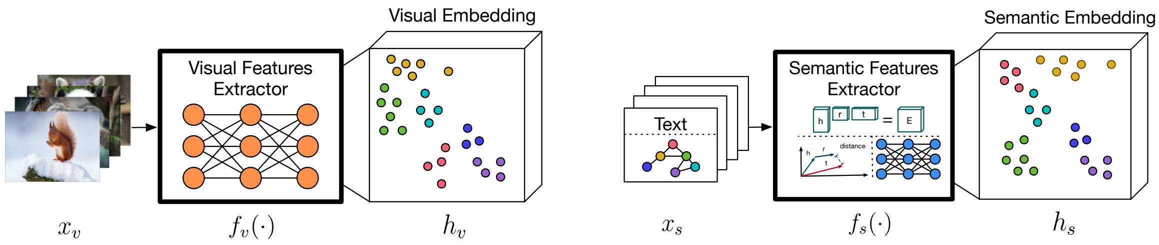
where the final dimensionality of $\mathbf{h}_s$ is determined by the architecture.

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 embeddings to hybrid approaches of vision and KGEs.

3.3. *Knowledge Graph Embedding*

A knowledge graph embedding method is part of the semantic features extractors, as shown in Section 3.2.2, and therefore a continuous vector representation $\mathbf{h}_{KG}$ of a discrete KG. Figure 3 illustrates a KGE-Method which transforms a KG into a KGE.

In the past, KGE-Methods were often decoupled from visual tasks and used more in the context of graph-based tasks such as node classification or link prediction. Due to the relationship to other semantic features extractors, such as language embeddings, we see great potential for KGE-Methods in visual object classification, detection, or segmentation. It is important to introduce several KGE-Methods and their categories since the final semantic embedding of a KG is strongly influenced by them.



(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.

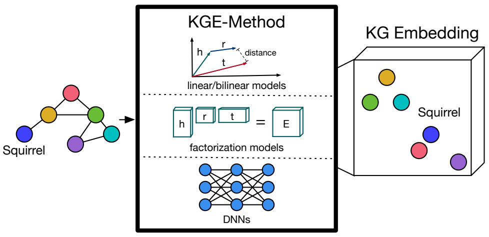


Fig. 3. A KGE-method transforms a KG into a KG Embedding.

3.3.1. *Unsupervised vs. Supervised KGE-Methods*

Concerning Ji et al. [16], every KGE-Method can be categorized based on the 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) that is used. However, for purpose of the survey, the categorization of graph embedding algorithms into unsupervised and supervised approaches proposed by Chami et al. [17] is more appropriate.

Unsupervised KGE-Methods only make use of the graph structure and its node features to form an embedding space. They do not consider task-specific labels for the graph or its nodes. Methods can be divided into shallow embedding methods, auto-encoders, and graph neural networks (GNNs). KGE-Methods of the shallow embedding methods learn a simple embedding lookup. These methods are transductive and therefore cannot be extended or transferred to other domains than the training domain. They can be either distance-based (TransE [18]), to force nodes that are close in the graph to be close in the embedding space or similarity-based, to remain similarities of the nodes using a dot-product (RESCAL [19]). For further insights, we refer to the survey of Wang et al. [20]. KGE-Methods using auto-encoders instead can encode non-linear complex structures of graphs, by using DNN encoder and decoder functions (SDNE [21]). KGE-Methods based on GNNs can use node features in addition to the graph structure. They pass information to neighbor nodes until some stable equilibrium state is reached (VGAE [22], DGI [23]).

Supervised KGE-Methods form the embedding space by using task-specific labels for the graph and its node features. Methods are categorized into shallow embedding methods, graph regularization methods, and graph convolution methods. Supervised shallow embedding methods learn an embedding lookup likewise to its unsupervised counterpart. However, their goal is to perform well on downstream tasks as link prediction or node classification, instead of learning a good graph representation only (LP [24]). Rather than learning an embedding lookup, KGE-Methods of the graph regularization methods learn the embedding as a parametric function defined over node features. This enables them to inductive settings, where a learned graph embedding is used on other domains. They can be further divided into Laplacian methods (ManiReg [25]) and skip-gram methods (node2vec [26], DeepWalk [27]). A subcategory of GNNs are graph convolutional networks (GCN [28]). Spectral graph convolutions apply convolutions in the spectral domain of the graph Laplacian matrix. Spectrum-based graph convolutional methods are limited by their domain dependency and cannot be applied in inductive settings (SCNN [29]), and spectrum-free methods require storing the entire graph adjacency matrix, which can be computationally expensive for large graphs (GCN [28]). Spatial graph convolutional methods use ideas such as neighborhood sampling and attention mechanisms to overcome challenges posed by graph irregularities (GAT [30]) None Euclidian graph convolutional methods yield significant improvements on graphs with hierarchical structure.

3.3.2. Hyper-relational Graph and Hypergraph KGE-Methods

The majority of existing KGE-methods and therefore most of the applications in the visual domain only work with directed labeled graphs, expecting binary relations in a tripled-based format. However, as shown in Section 3.1, a basic triplet representation oversimplifies the complex nature of the information that can be stored in hyper-relational graphs and hypergraphs [31]. Therefore, we provide a broad overview of the possibilities of transforming more complex KGs into appropriate KGEs so that they can be used in combination with DNNs. It is always possible to transform higher-relational graphs or hypergraphs into directed labeled graphs, so that standard KGE-Methods can be used. *Reification* converts more complex graphs into binary-relation graphs, by creating additional triplets from a hyper-relational fact and *star-to-clique* converts a tuple defined on k entities into $\binom{k}{2}$ tuples [32]. However, these conversions lead to suboptimal and incomplete models, since they only convert a set of key-value pairs, that are unaware of the triplet structure [31, 32]. Therefore, recently standard KGE-Methods have been adopted to directly operate on hyper-relational graphs (m-TransH [33], HypE, HSimple [32], RAE [34], GETD[35], TuckER [36], NaLP[37], HINGE[31], StarE [38]) and hypergraphs (HEBE [39], HGE [40], Hyper2vec [41], HNN [42], HCN [43], DHNE [44], HHNE [45], Hyper-SAGNN [46], HypE [32]).

3.4. Training Objectives for Joint Embeddings

Since visual and semantic information can be encoded in a vector-based embedding space, there are several training objectives to learn a joint representation. These objectives and also the DNNs are optimized mainly using *stochastic gradient descent* (SGD). 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 [1]. 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 [47, 48] showed that the objectives have a smaller impact on learned DNN than suspected, there are configurations of visual and semantic embedding space that only allow certain objectives to be applied. We define $\mathbf{l} \in \mathbb{R}^K$ as the network's output ("logit") vector, and $\mathbf{t} \in 0, 1^K$ 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.

3.4.1. Pointwise Objectives

Softmax Cross-Entropy (CE) [49] 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 $\mathbf{l}$ with the target vector $\mathbf{t}$ and is given by

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

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

Mean Squared Error (MSE) is the most intuitive way of attracting two vectors is using the MSE given by

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

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 [50].

However, using the Euclidian distance as a metric fails in high-dimensional space [51]. A more appropriate metric in high dimensions is the cosine distance given by $sim(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v} / \|\mathbf{u}\| \|\mathbf{v}\|$.

3.4.2. Pairwise Objectives

Pairwise objectives [52] always rely on the information of positive and negative samples. They have the goal to pull positive visual embedding vectors $\mathbf{h}_{v,p}$ to its corresponding semantic embedding anchor vector $\mathbf{h}_{s,a}$ and push negatives $\mathbf{h}_{v,n}$ away [53].

Triplet and Hinge Rank Loss [54] requires an explicit negative sampling. It uses a margin α as a regularization term and it is given by

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

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

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

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 is presented in [55] 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* and *Domain Generalization* are visual transfer learning tasks with labeled source data and unlabeled target data, where the former aims to extract implicit knowledge of classes in the source task T_S and transfers this knowledge to unknown classes of the target task T_T, and the latter aims to extract implicit knowledge of the source domain D_S and transfer this knowledge to an unknown target domain D_T. If both approaches additionally have access to a minimal set of labeled target data X_T, we call the task few-shot learning or domain adaptation.

Visual Transfer Learning using Knowledge Graphs has proven to be particularly advantageous compared to approaches without auxiliary knowledge [50, 56]. 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 [6]. 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 3.1 presents three distinct modeling structures with different levels of expressiveness and Section 3.3 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 us to make better predictions even when visual data is sparse or erroneous. However, there are several ways the auxiliary knowledge of a KG can be induces into the DL-Pipeline.

Referring to **RQ1**, this section provides a categorization of visual transfer learning approaches that combine KGs with the DL pipeline. As shown in Figure 4, we categorize the field of visual transfer learning using knowledge graphs into: 1) *Knowledge Graph as a Reviewer*, where the KG expects the visual embedding as input; 2) *Knowledge Graph as a Trainee*, where the KG uses the visual embedding as an objective; 3) *Knowledge Graph as a Trainer*, where the KG suits as an objective for the visual model; and 4) *Knowledge Graph as a Peer*, where the KG and the visual model are jointly used as an objective. Due to the shortage of visual transfer learning approaches for category 3) or 4) and their similarities to KGEs as explained in Section 3.2.2, we also consider approaches that use other semantic embedding spaces such as language embeddings as auxiliary knowledge.

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 task is given in Table 2.

4.1. Knowledge Graph as a Reviewer

Figure 5 shows the idea of using a KG as a reviewer. Visual model and KG are in a sequential order. The KG acts as a reviewer, using the independent output of a DNN as input to a graph or graph-based network to enrich the final prediction with auxiliary knowledge. If the model is learned end-to-end, the weights of the KG are optimized based on the visual input and the task, where the KG is used to reason over the output of a visual model. However, unlike the other categories, the KG as a reviewer does not align visual and semantic embedding space.

Most of the approaches map the output of a visual features extractor on the corresponding input nodes in a hierarchical graph, to enrich the output with interclass relationships. Lampert et al. [57] train a *support vector machine* (SVM) on SIFT features to predict binary *animals with attributes* (AwA) dataset attributes.

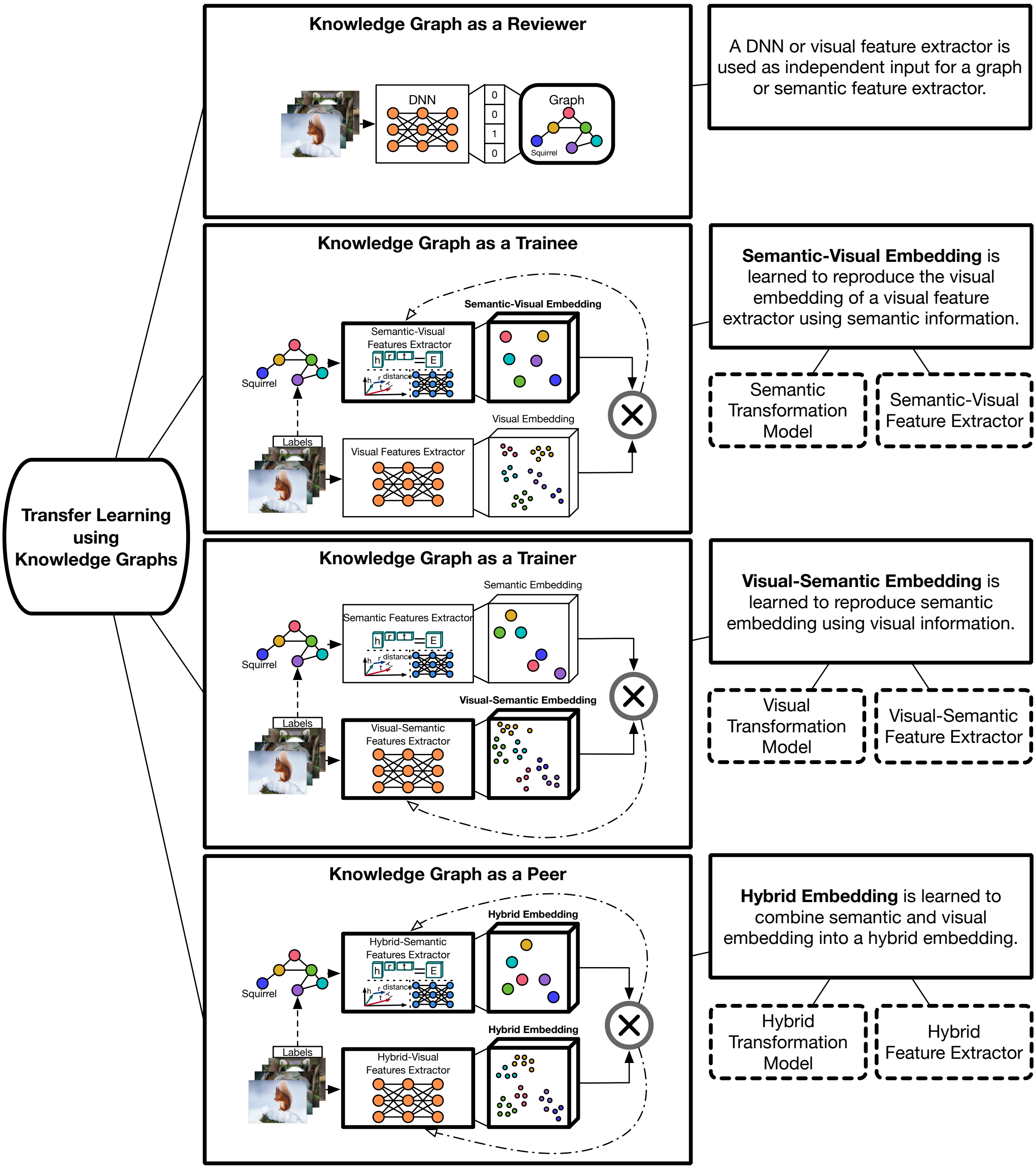

Transfer Learning using Knowledge Graphs
Knowledge Graph as a Reviewer
DNN
0
0
1
0
Graph
Squirrel
A DNN or visual feature extractor is used as independent input for a graph or semantic feature extractor.
Knowledge Graph as a Trainee
Semantic-Visual Embedding
Semantic-Visual Features Extractor
h r t =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

These class attributes are fed into a hierarchical graph-based network to predict unknown classes for a zero-shot learning task. Salakhutdinov et al. [58] introduce a hierarchical Bayesian classification model [59] 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

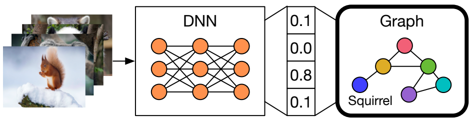


Fig. 5. Approaches from the category *Knowledge Graph as a Reviewer* use the KG for post validation on an independent DNN or visual features extractor.

method using a fixed graph based on WordNet[7] [60] and other approaches without hierarchical graph information. Deng et al. [61] proposed the *DARTS* algorithm for zero-shot learning. They pre-train an SVM on SIFT features of the ImageNet [62] dataset and map its classification output to WordNet with a reward and an accuracy to maximize the information gain. Ordonez et al. [63] 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. [64] 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. [65] introduce a *hierarchy and exclusion* (HEX) graph implemented as a network layer that introduces hierarchical concepts. For the HEX 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 using HEX graphs and evaluated their approach on ImageNet, in object classification, and zero-shot learning. Gebru et al. [66] use WordNet attributes to improve fine-grained object classification on the task of domain generalization with the office [67] and the large-scale cars dataset [68]. 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 instead of using CE or hard constraints as the HEX graphs [65]. Lee et al. [69] 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 [70] 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. [72] improves fine-grained image classification by creating a KG using the most common object-attribute and object-object relationships of the Visual Genome [105] dataset and higher-level semantics from WordNet. The output of a pre-trained, faster R-CNN [106] 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. [73] 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 [107] and Visual Genome, to connect classes with semantically related classes. Jiang et al. [74] extend [73] with *hybrid knowledge routed modules* (HKRM) by allowing them to be applied to intermediate feature representations and checking 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. [75] improve object detection by feeding the final object detections into a GCN which is based on object relationships and learned from MSCOCO dataset [108]. Gong et al. [71] propose a human pars-

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

Category	Sub-Category	Task Transfer	Domain Transfer	Other
Knowledge Graph as a Reviewer		[57], [61], [64], [65]	[66], [71]	[58], [63], [72], [73], [74], [75], [76]
Knowledge Graph as a Trainee	Semantic-Visual Transformation Model	[77], [78]		
	Semantic-Visual Features Extractor	[56], [79], [80], [81], [82]		[83]
Knowledge Graph as a Trainer	Visual-Semantic Transformation Model	[84], [50], [53], [85], [86], [87]		[88]
	Visual-Semantic Features Extractor		[89]	[90], [91]
Knowledge Graph as a Peer	Hybrid Transformation Model	[92], [93], [94], [95], [96], [97], [98]	[92]	[99], [100], [101], [102]
	Hybrid Features Extractor	[103]		[104]

Table 2

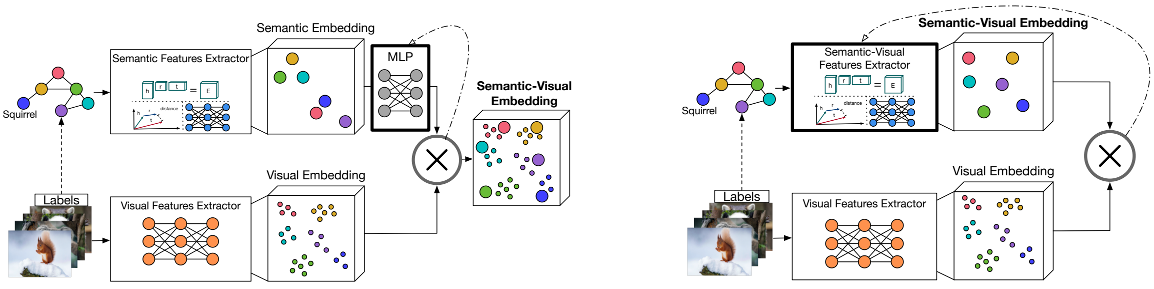
Categories and their tasks: Task transfer refers to the category zero and few-shot learning, domain transfer refers to the category domain generalization and adaptation, and other relates to object classification, object detection and object segmentation on source task and domain only.

ing 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. [76] present a *symbolic graph reasoning* (SGR) layer for semantic segmentation and image classification. It consists of a module that assigns the class features to the corresponding nodes of the KG, a graph reasoning over all previously defined nodes, and a mapping from the symbolic graph information back to the vector space. Their graph is based on an object relation graph from Visual Genome and a hierarchical relation graph from WordNet.

4.2. Knowledge Graph as a Trainee

As illustrated in Figure 6 visual and semantic model are organized in a parallel order. Approaches that belong to the category *Knowledge Graph as a Trainee* leverage auxiliary knowledge by providing a structure for a semantic model, e.g. GNN, that is learned using a visual embedding and rely on the idea that semantic similar classes should also have similar visual embedding vectors. Unlike the *Knowledge Graph as a Reviewer*, which uses the visual embedding as input for the KG, approaches from the category *Knowledge Graph as a Trainee* use the visual embedding as an objective. The KG acts as a trainee and optimizes its semantic embedding using the supervision of a visual embedding to a semantic-visual embedding. To combine visual and semantic information, some approaches either learn a transformation function, e.g. MLP, on a pre-trained semantic embedding space, e.g. language embedding, or apply GNNs to learn a semantic-visual features extractor under supervision of a visual embedding. Therefore, the fixed visual relationship between classes is used to learn the weights of the semantic relationship between classes. The advantage of a KG-based semantic-visual feature extractor is that relationships to other classes can be explicitly defined after training, so that new classes can be added without having to retrain the embedding method. This is enabled by the inductive property of GNNs, and is mostly used in zero-shot learning tasks. However, most of the approaches use a combination of fixed and dynamic semantic embeddings by initializing the nodes of the GNN using a language embedding and learning the weights of the GNN using the output of the visual embedding of the visual features extractor. Nevertheless, we count these methods as semantic-visual features extractors.

Semantic-Visual Transformation Models: As shown in Figure 6a, the pre-trained semantic features extractor is fixed over the whole training process and an additional transformation function, e.g. MLP, is learned to transform semantic information, e.g. class label names, into the semantic-visual embedding space. Approaches that use language models as fixed embeddings, could be replaced by KGEs, e.g. using shallow embedding


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 that learns the final embedding directly.

methods as defined in Section3.3. We claim that the lack of methods using KG embeddings is precisely due to their novelty, and KGEs could further improve the following approaches.

Rochan et al. [77] 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. [78] extends the idea and suggests using the visual space, instead of the semantic space, as the main embedding space, to reduce the hubness problem that occurs in high dimensions.

Semantic-visual features extractors: As illustrated in Figure 6b the semantic-visual features extractor directly learns the semantic-visual embedding using the supervision of the visual embedding space.

Instead of a fixed semantic embedding, approaches belonging to this category use KGE-Methods that can adapt their embedding to the task, here the visual embedding space. Therefore, as defined in Section 3.3, most approaches use supervised embedding methods, as GNNs. Wang et al [56] 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. [83] 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. [79] 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. [80] 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. [81] propose a *knowledge transfer network* (KTN), which extends [56] 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. [82] present the *knowl-*

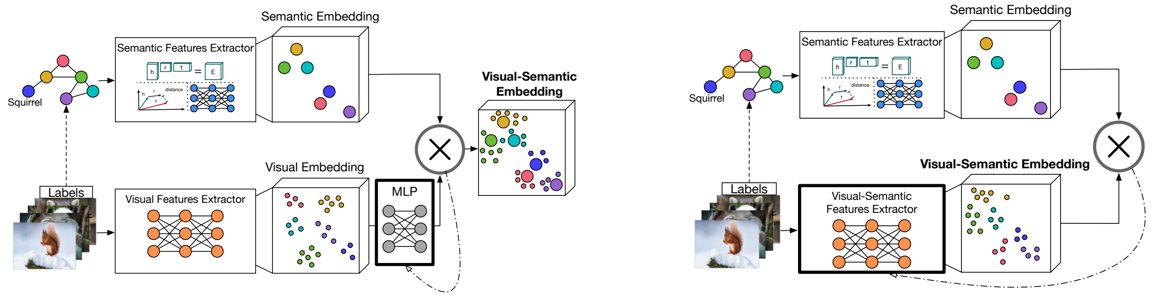
edge graph transfer network (KGTN). The knowledge graph transfer module incorporates a GGNN, which supports knowledge transfer of classes through a KG. To train GGNN, they fix the weights of a pre-trained visual features extractor and examine three different similarity metrics, such as inner product, 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.

4.3. Knowledge Graph as a Trainer

As shown in Figure 7 visual and semantic model are organized in a parallel order. Methods belonging to the *Knowledge Graph as a Trainer* category leverage auxiliary knowledge by influencing DNNs in learning specific visual features. Therefore, the KG acts as a trainer and supervises the training of the DNN using its semantic embedding space, rather than letting the DNN learn an independent visual embedding space based on the data distribution of the images. We refer to such an embedding of visual information learned under the supervision of a semantic embedding as a visual-semantic embedding. To combine semantic and visual information, some approaches either learn a transformation function, e.g. MLP, on a pre-trained and fixed visual embedding space, e.g., language embedding, or apply pairwise loss functions to learn a visual-semantic features extractor and thus a visual-semantic embedding space. Therefore, the fixed semantic relationship between classes is used to learn the weights of the visual relationship between classes. The advantage of a visual-semantic feature extractor is that the semantic embedding also influences the type of learned visual features extracted from the dataset, which can be helpful in scenarios with changing domains. Although, most of the approaches use language models as fixed embeddings, KGEs could be applied straightforwardly, e.g. using shallow embedding methods as defined in Section 3.3. We claim that the lack of methods using KG embeddings is precisely due to their novelty, and KGEs could further improve the following approaches.

Visual-Semantic Transformation Models are learned via a transformation function, e.g. MLP, from a pre-trained visual embedding space into the semantic embedding space, as depicted in Figure 7a. One of the first approaches that use semantic embeddings with NNs is the work from Mitchell et al. [88]. They use language embeddings derived from text corpus statistics to generate neural activity pattern images. Instead of generating images from text, Palatucci et al. [84] learn a linear regression model to map neural activity patterns into language embedding space. Socher et al. [50] 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. [53] 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 [109]. In contrast to Socher et al. [50] 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. [85] 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. [86] 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. Akata et al. [87] 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 [109] label embedding space.

Visual-semantic features extractors: The approaches mentioned so far only learn a transformation from visual embedding to semantic embedding. However, the parameters of the feature extractors are not affected by the auxiliary information. Thus, if the feature extractor cannot detect visual features due to the domain shift problem, the performance of the final prediction suffers. A conceptual architecture is depicted in Figure 7b


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

(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.

where the weights of the feature extractor are directly influenced by the semantic embedding space.

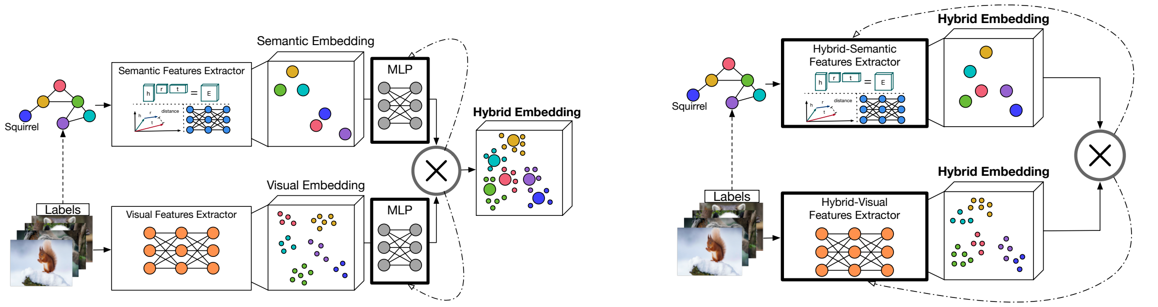
Joulin et al. [90] demonstrate that feature extractors trained to predict words in image captions learn useful image representations. They converted the title, description, and hashtag metadata of images into a bag-of-words multi-label classification task and showed that pre-training a feature extractor to predict these labels learned representations which performed similarly to ImageNet-based pre-training on transfer tasks. Radford et al. [91] 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. [104]. Recently, Monka et. al [89] presented an approach for the task of domain adaptation that uses a directed labeled KG of road signs (RoadSignKG). To use the KG in combination with a DNN, the KG is transformed into a vector-based KGE. They propose a contrastive learning method that is supervised by the KGE to learn the weights of the visual features extractor. They show that their visual features extractor learned using the *Knowledge Graph as a Trainer* outperforms a conventional DNN trained with CE and a similar DNN without auxiliary information of the KG in visual transfer learning tasks.

4.4. *Knowledge Graph as a Peer*

As depicted in Figure 8 visual and semantic model are organized in a parallel order. Approaches of the category *Knowledge Graph as a Peer* leverage auxiliary knowledge by influencing semantic and visual embedding equally. Unlike previous approaches, the idea of a hybrid embedding space is to fuse the visual embedding and semantic embedding to a hybrid embedding space that contains information of both spaces. The final hybrid embedding space is either a combination of pre-trained visual and semantic embedding, learned by a transformation function, e.g. MLP, or a combination of hybrid-visual and hybrid-semantic features extractors, trained using a joint loss function.

Hybrid Transformation Models are learned via a transformation function from pre-trained visual and semantic embeddings into the hybrid embedding space. As illustrated in Figure 8a semantic and visual features extractors are fixed over the whole training process and additional transformation functions, e.g. MLPs, are learned to combine both spaces in a hybrid embedding space.

Yang et al. [92] 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. [93] 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 learn-



(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.

ing in the transductive embedding space, which combines additional semantic knowledge in the form of attributes and word vectors from related classes. Ba et al. [94] 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. [99] 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. [95] 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. [96] 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 [110], Word2Vec attributes, and GloVe attributes. Zhao et al. [97] 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. Tang et al. [100] 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. [98] 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. [101] propose a multi-layer transformer [111] 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. [102] 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.

Hybrid Features Extractors are randomly initialized and trained by supervision of the hybrid embedding space. Thus there is no additional transformation function needed, as a hybrid-visual features extractor and a hybrid-semantic features extractor use the same hybrid embedding for visual and semantic input data, as depicted in Figure 8b.

Zhang et al. [104] 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. Recently, Naeem et. al [103] 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.

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 about 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) Zero-shot datasets without auxiliary knowledge, used to prove the ability of an approach to transfer to novel classes. e) Domain generalization datasets without auxiliary knowledge, used to prove the ability of an approach to transfer to novel domains.

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. [112].

WordNet [60]: 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 [113]: 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 3.1), where concepts are connected via binary relationships. It contains approximately 34 million statements, i.e. edges [8].

DBPedia [114]: 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

[8] https://conceptnet.io

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

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.

on the Web to Wikipedia data. The underlying structure of DBpedia is a hypergraph model (see Section 3.1) 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 [115]: 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 [116] 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 3.1) where relations are enriched with additional information known as qualifiers [38]. 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. 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.

AwA [57]: 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. We hope that this dataset will facilitate research and serve as a testbed for attribute-based classification. However, as AWA images does not have the public copyright license, only some image features, i.e. SIFT [14], DECAF [117], VGG19 [118] of AWA

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

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

dataset is 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 [119]: 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 [120]: 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. [87] introduces the first zero-shot split of CUB with 150 training, 50 validation, and 50 test classes.

SUN [121]: 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 14340 images coming from 717 types of scenes annotated with 102 attributes. Lampert et al. [110] use 645 classes of SUN for training, 65 classes for validation, and 72 classes for testing.

5.2.2. Language Augmented Image Datasets

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

MS-COCO [108]: 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 [122] released an additionally learned denotation graph for MS-COCO, which induces a partial ordering over the textual image descriptions.

Flickr30K [123]: 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 [122].

SBU Captions [124] 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] 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 [105] 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.

ImageNet [126]: 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.

miniImageNet [127] 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.

tiredImageNet [128] 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. These are split 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.

5.2.4. Zero-Shot Datasets without Auxiliary Knowledge

This section introduces datasets without auxiliary knowledge that have been artificially extended with KGs or auxiliary knowledge in recent works. However, they were originally created without additional knowledge.

CIFAR-10 [129] is an object recognition dataset with 10 classes. The dataset contains 80 mil color images downscaled to 32 × 32 and spread out across 79K search terms. In addition, it provides an unsupervised learning subset of about 2 mil unlabeled images.

CIFAR-100 [129]: contains 100 classes for object recognition and has the same properties as the CIFAR-10 dataset. It contains 600 images in each of 100 classes. The 100 classes are further grouped into 20 superclasses.

CIFAR-FS [130] is randomly sampled from CIFAR-100 by using the same criteria with which miniImageNet has been generated. Moreover, 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 [131]: 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 the 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.

5.2.5. Domain Generalization Datasets without Auxiliary Knowledge

In addition, we introduce widely used domain generalization or domain adaptation datasets without auxiliary knowledge.

Office-31 [67] 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 [132]: 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 [133]: 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. [134] and Zhang et al. [135] 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. [136] 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. [137] reviewed and structured 20 transfer-learning approaches. Wang et al. [138] 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. [139] 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. [135] 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. [140] 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. [119] 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. Aditya et al. [141] investigated representative reasoning mechanisms, knowledge integration methods, and their corresponding image understanding applications. Therefore, they divide auxiliary knowledge used in CV into knowledge about objects, regions, and actions, and high-level common-sense knowledge and provide an overview about frameworks that are capable of logic operations. Further, they briefly discuss the knowledge integration in the DL era and categorization approaches into: i) pre-process domain knowledge and augment training samples, ii) vectorize parts of knowledge base and input to intermediate layers, iii) inspire neural network architecture from an underlying knowledge graph, iv) post-process and reason with external knowledge. We also include ii) and iv) in our category *Knowledge Graph as a Reviewer* since we see knowledge layers in the DNN as an intermediate reviewing and validation process. Category iii) focuses on knowledge-based teacher models that inspire a student DNN using knowledge distillation. We can see similarities of iii) to the category *Knowledge Graph as a Trainer*. However, we take this one step further and provide additional categories of combination, based on which type of information inspires which. The *Knowledge Graph as a Trainer*, where the KG inspires the visual DNN, the *Knowledge Graph as a Trainee* where the KG gets inspired by the visual DNN, and the *Knowledge Graph as a Peer* where both KG and visual DNN inspire each other. Von Rueden et al. [142] 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. [143] 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. [144] 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. [145] investigated the field of explainable AI using KGs and categorized approaches into knowledge matching, cross-disciplinary and interactive explanations. Chen et al. [146] and Chari et al. [147] 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. [148] summarized semantic web technologies that can provide valid explanations for ML models, separating them due to their ML technique and semantic expressiveness.

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 embedding spaces, and different methods for converting KGs into a vector-based KGE. The main contribution is a categorization into four individual ways of how a KG can be used with a DL pipeline for visual transfer learning tasks. To enable fair comparisons for approaches in the field 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, 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 3.1, 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. [149–151]. 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 3.2.2, as a semantic features extractor either can be used a knowledge graph embedding model or a language model to form the respective embedding spaces. With this assumption, we can apply KGEs in various new domains, where language embeddings have shown a potential for improving the robustness, with the advantage that the KGE space 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 KGE 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 3.4, 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 the KGE is influenced by the visual embedding;
  3) *Knowledge Graph as a Trainer* - where the KGE influences the visual embedding; and
  4) *Knowledge Graph as a Peer* - where the KGE and the visual embedding influence each other.
- **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 the GNN enables reasoning over the output 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 semantic model, e.g. GNN, that is learned using a visual embedding. 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 a image data distribution independent embedding provided by a semantic embedding instead of just using the data distribution. Thus, we see the advantage of these approaches specifically in domain generalization scenarios.
  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, and SUN.
  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.

Zero-Shot Datasets without Auxiliary Knowledge, as CIFAR-10, CIFAR-100, CIFAR-FS, and FC100.

Domain Generalization Datasets without Auxiliary Knowledge, 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.

9. Acknowledgement

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “KI Delta Learning – Development of methods and tools for the efficient expansion and transformation of existing AI modules of autonomous vehicles to new domains. The authors would like to thank the consortium for the successful cooperation.”

References

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

[2] 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, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III*, V. Kurková, Y. Manolopoulos, B. Hammer, L.S. Iliadis and I. Maglogiannis, eds, Lecture Notes in Computer Science, Vol. 11141, Springer, 2018, pp. 270–279. doi:10.1007/978-3-030-01424-7_27.

[3] 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, Conference Track Proceedings*, Y. Bengio and Y. LeCun, eds, 2015. http://arxiv.org/abs/1412.6572.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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* **abs/2003.02320** (2020). https://arxiv.org/abs/2003.02320.

[11] 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.

[12] 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.

[13] T. Chen, S. Kornblith, K. Swersky, M. Norouzi and G.E. Hinton, Big Self-Supervised Models are Strong Semi-Supervised Learners, in: *NeurIPS*, 2020.

[14] 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.

[15] 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), 20-26 June 2005, San Diego, CA, USA*, IEEE Computer Society, 2005, pp. 886–893. doi:10.1109/CVPR.2005.177.

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

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

[18] A. Bordes, N. Usunier, A. García-Durán, J. Weston and O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 2787–2795. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html.

[19] M. Nickel, V. Tresp and H. Kriegel, A Three-Way Model for Collective Learning on Multi-Relational Data, in: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, L. Getoor and T. Scheffer, eds, Omnipress, 2011, pp. 809–816. https://icml.cc/2011/papers/438_icmlpaper.pdf.

[20] 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.

[21] D. Wang, P. Cui and W. Zhu, Structural Deep Network Embedding, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, ACM, 2016.

[22] T.N. Kipf and M. Welling, Variational Graph Auto-Encoders, *CoRR* **abs/1611.07308** (2016). http://arxiv.org/abs/1611.07308.

[23] P. Velickovic, W. Fedus, W.L. Hamilton, P. Liò, Y. Bengio and R.D. Hjelm, Deep Graph Infomax, 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=rklz9iAcKQ.

[24] X. Zhu and Z. Ghahramani, Learning from labeled and unlabeled data with label propagation, 2002.

[25] M. Belkin and P. Niyogi, Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, in: *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, MIT Press, 2001.

[26] A. Grover and J. Leskovec, node2vec: Scalable Feature Learning for Networks, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, ACM, 2016.

[27] B. Perozzi, R. Al-Rfou and S. Skiena, DeepWalk: online learning of social representations, in: *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, ACM, 2014.

[28] 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.

[29] J. Bruna, W. Zaremba, A. Szlam and Y. LeCun, Spectral Networks and Locally Connected Networks on Graphs, in: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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, December 12-15, 2016, Barcelona, Spain*, 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.

[40] 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.

[41] 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, Chiang Mai, Thailand, April 22-25, 2019, Proceedings, Part III, and DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA, Chiang Mai, Thailand, April 22-25, 2019, Proceedings*, 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.

[42] 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.

[43] N. Yadati, M. Nimishakavi, P. Yadav, V. Nitin, A. Louis and P.P. Talukdar, 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, December 8-14, 2019, Vancouver, BC, Canada*, 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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, Proceedings, Part VI*, 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.

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

[49] 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.

[50] 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 December 5-8, 2013, Lake Tahoe, Nevada, United States*, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 935–943.

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

[52] 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), 17-22 June 2006, New York, NY, USA*, IEEE Computer Society, 2006, pp. 1735–1742. doi:10.1109/CVPR.2006.100.

[53] 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 December 5-8, 2013, Lake Tahoe, Nevada, United States*, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 2121–2129.

[54] 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.

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

[56] 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. doi:10.1109/CVPR.2018.00717. http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_Zero-Shot_Recognition_via_CVPR_2018_paper.html.

[57] 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), 20-25 June 2009, Miami, Florida, USA*, IEEE Computer Society, 2009, pp. 951–958. doi:10.1109/CVPR.2009.5206594.

[58] 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.

[59] B. Shahbaba and R. Neal, Improving Classification When a Class Hierarchy is Available Using a Hierarchy-Based Prior, *Bayesian Analysis* **2** (2005), 221–237.

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

[61] 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.

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

[63] 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.

[64] 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 December 5-8, 2013, Lake Tahoe, Nevada, United States*, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 46–54.

[65] 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, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, 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.

[66] 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.

[67] 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, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV*, 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.

[68] 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, February 4-9, 2017, San Francisco, California, USA*, S.P. Singh and S. Markovitch, eds, AAAI Press, 2017, pp. 4502–4508. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14583.

[69] 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. doi:10.1109/CVPR.2018.00170. http://openaccess.thecvf.com/content_cvpr_2018/html/Lee_Multi-Label_Zero-Shot_Learning_CVPR_2018_paper.html.

[70] 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, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, A. Moschitti, B. Pang and W. Daelemans, eds, ACL, 2014, pp. 1532–1543. doi:10.3115/v1/d14-1162.

[71] 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. doi:10.1109/CVPR.2019.00763. http://openaccess.thecvf.com/content_CVPR_2019/html/Gong_Graphonomy_Universal_Human_Parsing_via_Graph_Transfer_Learning_CVPR_2019_paper.html.

[72] 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.

[73] 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. doi:10.1109/CVPR.2018.00756. http://openaccess.thecvf.com/content_cvpr_2018/html/Chen_Iterative_Visual_Reasoning_CVPR_2018_paper.html.

[74] 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 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, 2018, pp. 1559–1570.

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

[76] 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, December 3-8, 2018, Montréal, Canada*, S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, 2018, pp. 1858–1868.

[77] 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.

[78] 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.

[79] 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. doi:10.1109/CVPR.2019.01175. http://openaccess.thecvf.com/content_CVPR_2019/html/Kampffmeyer_Rethinking_Knowledge_Graph_Propagation_for_Zero-Shot_Learning_CVPR_2019_paper.html.

[80] 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.

[81] 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.

[82] 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.

[83] 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. doi:10.1109/CVPR.2019.00532. http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_Multi-Label_Image_Recognition_With_Graph_Convolutional_Networks_CVPR_2019_paper.html.

[84] 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 7-10 December 2009, Vancouver, British Columbia, Canada*, Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams and A. Culotta, eds, Curran Associates, Inc., 2009, pp. 1410–1418.

[85] 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, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, eds, 2014. http://arxiv.org/abs/1312.5650.

[86] 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.

[87] 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.

[88] 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. doi:10.1126/science.1152876. https://science.sciencemag.org/content/320/5880/1191.

[89] S. Monka, L. Halilaj, S. Schmid and A. Rettinger, ConTraKG: Contrastive-based Transfer Learning for Visual Object Recognition using Knowledge Graphs (2020).

[90] 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, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, 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.

[91] 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.

[92] 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, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, eds, 2015. http://arxiv.org/abs/1412.7489.

[93] 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.

[94] 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, De-*

cember 7-13, 2015, IEEE Computer Society, 2015, pp. 4247–4255. doi:10.1109/ICCV.2015.483.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, 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.

[102] 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* **abs/2006.16934** (2020). https://arxiv.org/abs/2006.16934.

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

[104] 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* **abs/2010.00747** (2020). https://arxiv.org/abs/2010.00747.

[105] 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.

[106] 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.

[107] 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.

[108] 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, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 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.

[109] 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 December 5-8, 2013, Lake Tahoe, Nevada, United States*, C.J.C. Burges, L. Bottou, Z. Ghahramani and K.Q. Weinberger, eds, 2013, pp. 3111–3119.

[110] 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.

[111] 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, December 4-9, 2017, Long Beach, CA, USA*, 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.

[112] 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.

[113] 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, February 4-9, 2017, San Francisco, California, USA*, S.P. Singh and S. Markovitch, eds, AAAI Press, 2017, pp. 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972.

[114] 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.

[115] D. Vrandecic, Wikidata: a new platform for collaborative data collection, in: *Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20,*

2012 (Companion Volume), ACM, 2012, pp. 1063–1064. doi:10.1145/2187980.2188242.

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

[117] 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.

[118] K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, in: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, eds, 2015. http://arxiv.org/abs/1409.1556.

[119] 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.

[120] 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.

[121] 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, 2012, pp. 2751–2758. doi:10.1109/CVPR.2012.6247998.

[122] 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.

[123] 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.

[124] 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 12-14 December 2011, Granada, Spain*, 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.

[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, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, I. Gurevych and Y. Miyao, eds, Association for Computational Linguistics, 2018, pp. 2556–2565. doi:10.18653/v1/P18-1238. https://www.aclweb.org/anthology/P18-1238/.

[126] 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.

[127] 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, December 5-10, 2016, Barcelona, Spain*, D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon and R. Garnett, eds, 2016, pp. 3630–3638.

[128] 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, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, OpenReview.net, 2018. https://openreview.net/forum?id=HJcSzz-CZ.

[129] A. Krizhevsky, G. Hinton et al., Learning multiple layers of features from tiny images (2009).

[130] 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.

[131] 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, December 3-8, 2018, Montréal, Canada*, S. Bengio, H.M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds, 2018, pp. 719–729.

[132] 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.

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

[134] 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.

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

[136] 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.

[137] M. Kaboli, A Review of Transfer LearningAlgorithms, Research Report, Technische Universität München, 2017, Transfer Learning Algorithms. https://hal.archives-ouvertes.fr/hal-01575126.

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

[139] 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.

[140] 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.

[141] 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.

[142] 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* **abs/1903.12394** (2019). http://arxiv.org/abs/1903.12394.

[143] 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, Palo Alto, CA, USA, March 23-25, 2020, Volume I*, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle and F. van Harmelen, eds, CEUR Workshop Proceedings, Vol. 2600, CEUR-WS.org, 2020. http://ceur-ws.org/Vol-2600/paper7.pdf.

[144] 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.

[145] 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.

[146] 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.

[147] 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.

[148] 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.

[149] 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.

[150] 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.

[151] 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.