

# Multi-Modal Knowledge Graph Construction and Application: A Survey

Xiangru Zhu, Zhixu Li *Member, IEEE*, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao *Member, IEEE*, Nicholas Jing Yuan *Member, IEEE*,

**Abstract**—Recent years have witnessed the resurgence of knowledge engineering which is featured by the fast growth of knowledge graphs. However, most of existing knowledge graphs are represented with pure symbols, which hurts the machine’s capability to understand the real world. The multi-modalization of knowledge graphs is an inevitable key step towards the realization of human-level machine intelligence. The results of this endeavor are Multi-modal Knowledge Graphs (MMKGs). In this survey on MMKGs constructed by texts and images, we first give definitions of MMKGs, followed with the preliminaries on multi-modal tasks and techniques. We then systematically review the challenges, progresses and opportunities on the construction and application of MMKGs respectively, with detailed analyses of the strength and weakness of different solutions. We finalize this survey with open research problems relevant to MMKGs.

**Index Terms**—Multimodal Knowledge Graph, Survey, Symbol Grounding

## 1 INTRODUCTION

Recent years have witnessed the resurgence of knowledge engineering which is featured by the fast growth of knowledge graphs. A knowledge graph (KG) is essentially a large-scale semantic network that contains entities, concepts as nodes and various semantic relationships among them as edges. Knowledge graphs have found their great values in a wide range of real world applications including text understanding, recommendation system and natural language question answering. More and more knowledge graphs have been created, covering common sense knowledge (e.g. Cyc [1], ConceptNet [2]), lexical knowledge (e.g. WordNet [3], BabelNet [4]), encyclopedia knowledge (e.g. Freebase [5], DBpedia [6], YAGO [7], WikiData [8], CN-Dbpedia [9]), taxonomic knowledge (e.g. Probase [10]) and geographic knowledge (e.g. GeoNames [11]).

However, most of the existing knowledge graphs are represented with pure symbols denoted in the form of text, which weakens the capability of machines to describe and understand the real world. A human being cannot understand what a dog is without the experience of living with dog, which enlightens researchers to establish the connection between the symbol Dog and the experience of dogs, that is, grounding a symbol to its physical world meaning [12], [13], [14]. Similarly, grounding symbolic forms to non-symbolic experiences benefits to receive real communicative intents [15]. For example, the customers can-

not understand the meaning of Hand-in-waistcoat as a special pose (hand inside coat flap) without *experiences* of Hand-in-waistcoat so that the customer would give a wrong response to the request of photographers. Thus, it is necessary to ground symbols to corresponding images, sound and video data and map symbols to their corresponding referents with meanings in the physical world, enabling machines to generate similar “*experiences*” like a real human [12] when they are confronted with a specific entity Hand-in-waistcoat or an abstract concept Dog. On the other hand, there is an increasing demand for the multi-modality of knowledge to break through the bottleneck of real world applications [16], [17], [18]. For instance, in relation extraction tasks, an additional image usually greatly improves the performance in the extraction of the attributes and relationships that are visually obvious but difficult to be recognized in symbols and text, such as *partOf* (e.g. *The keyboard and the screen are parts of a laptop.*) and *colorOf* (e.g. *A banana is usually yellow or yellowish-green but not blue*). In text generation tasks, if the machine has been empowered with the ability to recognize a specific entity in an image by the reference to a Multi-Modal KG (MMKG), the machine is possible to generate a more informative entity-level sentence (e.g. *Donald Trump is making a speech*) instead of a vague concept-level description (e.g. *A tall man with blond hair is making a speech*).

Due to the rapid growth of applications’ demand for multi-modal knowledge guidance, the multi-modalization of KGs and their applications are booming in recent years. But there still lacks a systematic review of the recent research progresses, challenges and opportunities in this emerging area. In this paper, we hope to fill the gap and systematically survey the recent research progress relevant to MMKG as follows: 1) **Construction**. The construction of MMKGs could be conducted in two opposite directions. One is from images to symbols, i.e., labeling images with symbols in KG; the other is from symbols to images, i.e., grounding

• X. Zhu, Z. Li, X. Wang, X. Jiang, P. Sun, X. Wang and Y. Xiao are with the School of Computer Science, Fudan University.  
E-mail: {xrzhu19, zhixuli, xiaodanwang20, xueyaojiang19, plsun20, xuwwang18, shawyh}@fudan.edu.cn. Z. Li and Y. Xiao are the corresponding authors.

• N.J. Yuan is with Huawei Cloud & AI, Hangzhou, Zhejiang, China.  
E-mail: nicholas.yuan@huawei.com

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

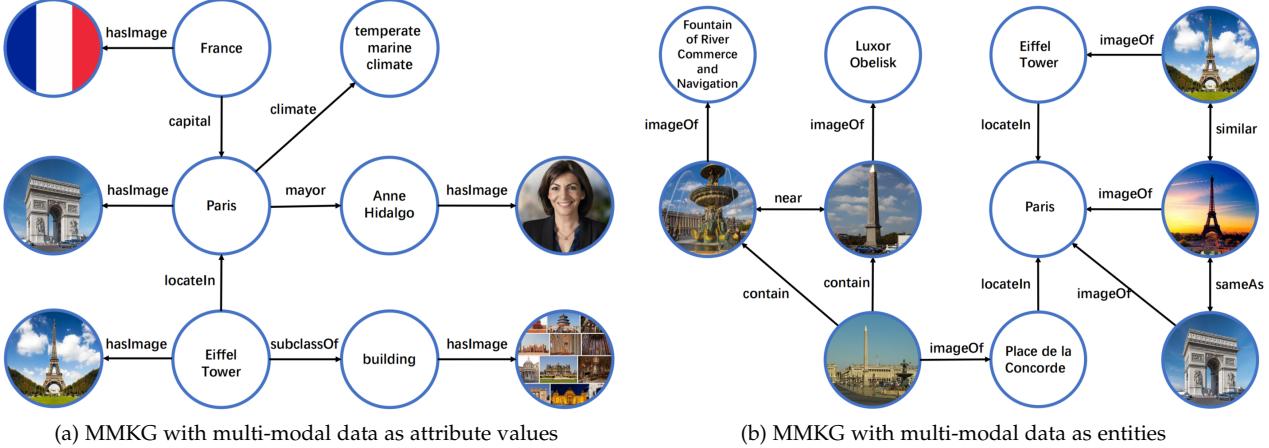


Fig. 1: Example MMKGs of two different types: A-MMKG and N-MMKG

symbols in KG to images. In the Construction section, we will systematically cover the challenges, progresses as well as opportunities on correlating various symbol knowledge (including entities, concepts, relations and events) to their corresponding images in the two opposite directions. 2) **Application.** The application of MMKGs could be roughly divided into two categories, which are In-MMKG applications aiming at addressing the quality or integration issues of MMKGs themselves, and Out-of-MMKG applications which are general multi-modal tasks that MMKGs can help. In the Application section, we will present how MMKGs are applied in several well-studied multi-modal tasks.

To summarize, we are the first to thoroughly survey the existing work on MMKGs consisting of texts and images. To enhance the value of this survey, we pay attention to ensure the following features: 1) **Comprehensive Survey.** We systematically and comprehensively review the existing work on MMKG construction and application. 2) **Insightful Analysis.** We analyse the strengthens and weakness of different solutions in MMKG construction, and also discuss on how MMKGs can help in various downstream applications. 3) **Revealed Opportunities.** We not only point out some potential opportunities with the studied tasks relevant to MMKG construction, but also list some promising future directions with MMKG.

The rest of the survey are organized as follows: Sec. 2 gives definitions and preliminaries on MMKGs. Sec. 3 conducts a comprehensive review on the challenges, progresses and opportunities on the construction of MMKGs, while Sec. 4 presents how MMKGs are applied in several well-studied multi-modal applications. Sec. 5 reviews some open problems of MMKG and highlight promising future directions. Sec. 6 finally concludes the paper.

## 2 DEFINITIONS AND PRELIMINARIES

This section first defines two representation ways for KGs, and then reviews some preliminaries on multi-modal tasks and techniques, followed with a discussion on the connections between MMKGs and the existing multi-modal tasks and techniques.

### 2.1 Definition & Representation of MMKGs

A traditional Knowledge Graph (KG) is defined as a directed graph  $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T}_{\mathcal{R}}, \mathcal{T}_{\mathcal{A}}\}$ , where  $\mathcal{E}$ ,  $\mathcal{R}$ ,  $\mathcal{A}$ ,  $\mathcal{V}$  are sets of entities, relations, attributes and literal attribute values, and  $\mathcal{T}_{\mathcal{R}} = \mathcal{E} \times \mathcal{R} \times \mathcal{E}$  and  $\mathcal{T}_{\mathcal{A}} = \mathcal{E} \times \mathcal{A} \times \mathcal{V}$  are sets of relation triples and attribute triples respectively. A triple  $(s, p, o) \in \mathcal{T}_{\mathcal{R}}$  denotes that entity  $s \in \mathcal{E}$  has relation  $p \in \mathcal{R}$  with entity  $o \in \mathcal{E}$ . A triple  $(s, p, o) \in \mathcal{T}_{\mathcal{A}}$  denotes that entity  $s \in \mathcal{E}$  has attribute  $p \in \mathcal{A}$  with the attribute value  $o \in \mathcal{V}$ .

A Multi-modal Knowledge Graph (MMKG) can be seen as a multi-modalized KG, which has part of its knowledge in  $\{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T}_{\mathcal{R}}, \mathcal{T}_{\mathcal{A}}\}$  multi-modalized. We say a particular knowledge symbol is multi-modalized if it is associated with its corresponding data items in modalities other than text, such as image, sound or video, that could embody the knowledge. For instance, a relation triple  $(s, p, o)$  can be multi-modalized with an image describing the relation  $p$  between  $s$  and  $o$ .

Existing work on MMKGs mainly adopts two different ways for representing MMKGs. One way takes multi-modal data (images in this survey) as particular attribute values of entities or concepts as the example shown in Fig. 1(a). We name a MMKG represented in this way as **A-MMKG** for short, denoted as  $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T}_{\mathcal{R}}, \mathcal{T}_{\mathcal{A}}\}$ , where  $\mathcal{T}_{\mathcal{A}} = \mathcal{E} \times \mathcal{A} \times (\mathcal{V}_{KG} \cup \mathcal{V}_{MM})$  is the set of attribute triples,  $\mathcal{V}_{KG}$  is the set of the KG's attribute values and  $\mathcal{V}_{MM}$  is the set of multi-modal data. In A-MMKGs, since multi-modal data are treated as attribute values, in a triple  $(s, p, o)$ ,  $s$  denotes an entity,  $o$  denotes one of its corresponding multi-modal data, and the relation  $p$  is "hasImage" when  $o$  is an image. Some example triplets are listed in Table 1a.

The other way takes multi-modal data as entities in KGs as the example shown in Figure 1(b). We name a MMKG represented in this way as **N-MMKG** for short, denoted as  $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T}_{\mathcal{R}}, \mathcal{T}_{\mathcal{A}}\}$ , where  $\mathcal{T}_{\mathcal{R}} = (\mathcal{E}_{KG} \cup \mathcal{E}_{MM}) \times \mathcal{R} \times (\mathcal{E}_{KG} \cup \mathcal{E}_{MM})$  is the set of relation triples,  $\mathcal{E}_{KG}$  is the set of KG entities and  $\mathcal{E}_{MM}$  is the set of multi-modal data. Since multi-modal data are treated as new entities, more inter-modal and intra-modal relations are discovered and added into the MMKG. For example, in Table 1b, the entity *Eiffel Tower.jpg* is associated with an image *Eiffel\_Tower.jpg* by the relation *imageOf*. Two images can also be associated in one of the following relations:

subject	predicate	object
France	hasImage	The_flag_of_France.jpg
Anne Hidalgo	hasImage	Anne_Hidalgo.jpg
Paris	mayor	Anne Hidalgo
Paris	hasImage	A_landmark_of_Paris.jpg
Eiffel Tower	locateIn	Paris
Eiffel Tower	hasImage	Eiffel_Tower.jpg
Eiffel Tower	subclassOf	building
building	hasImage	a_kind_of_architectural_style.jpg

(a) Example RDF triples in A-MMKG

subject	predicate	object
Eiffel_Tower_in_Paris.jpg	imageOf	Paris
Eiffel_Tower_in_Paris.jpg	size	700*1600
Eiffel_Tower_in_Paris.jpg	sameAs	Arc_de_Triomphe_in_Paris.jpg
Eiffel_Tower_in_Paris.jpg	similar	Eiffel_Tower.jpg
Eiffel Tower	locateIn	Paris
Eiffel_Tower.jpg	imageOf	Eiffel Tower
Eiffel_Tower.jpg.HOG	describes	Eiffel_Tower.jpg
Eiffel_Tower.jpg.HOG	value	[0.0775, 0.0120, 0.0021, ...]

(b) Example RDF triples in N-MMKG

TABLE 1: Example RDF triples in different types of MMKGs, where items end up with “.jpg” are images.

- **contain:** One image entity visually contains another image entity by the relative position of images.
- **nearBy:** One image entity is visually nearby another image entity in an image.
- **sameAs:** Two different image entities refer to the same entity.
- **similar:** Two image entities are visually similar to each other.

In addition, in N-MMKGs an image is usually abstracted into a number of image descriptors, which are usually summarized into feature vectors of the image entity at pixel level, such as Gray Histogram Descriptor (GHD), Histogram of Oriented Gradients Descriptor (HOG), Color Layout Descriptor (CLD) and so on. For example, in Table 1b, *Eiffel\_Tower\_in\_Paris.jpg.HOG* is one of descriptors of the image *Eiffel\_Tower\_in\_Paris.jpg*, and is in the form of a vector. These image descriptors are well interpreted, thus the relations between images can be obtained by simple calculations (e.g. image similarity obtained via inner product of vectors of image descriptors).

Table 2 lists the mainstream MMKGs constructed by texts and images with detailed information, including MMKG type, multi-modalized knowledge, source resources, candidate resources, quality control, scale, image per entity etc.

## 2.2 Preliminaries on Multi-Modal Techniques

The construction and application of MMKG has a great relation with the existing multi-modal researches in Computer Vision (CV). Here we briefly review some well-studied multi-modal tasks, multi-modal learning techniques, followed with important progresses on multi-modal pretrained language model.

**Multi-Modal Tasks.** Generally, modality refers to the particular way in which something exists, is experienced or is done [27]. In computer science and artificial intelligence, we say a problem is characterized as multi-modal if it involves data of multiple modalities. A multi-modal task integrates and models multiple communicative modalities, in order for acquiring knowledge or comprehension from the multi-modal data [27]. Some example multi-modal tasks based on images and texts are listed below:

- 1) **Image Captioning.** Image captioning aims at generating the descriptive caption for a given image [28], [29].
- 2) **Visual Grounding.** Visual grounding aims at locating an object with designated description in a given image [30], [31].

- 3) **Visual Question Answering (VQA).** VQA aims at generating a textual answer for a textual question with the help of a relevant image [32], [33].
- 4) **Cross-Modal Retrieval.** There are two classic cross-modal retrieval tasks including searching for images through a text [34], [35], and searching for texts through an image [36], [37].

**Multi-Modal Learning.** The multi-modal learning mainly focus on modeling the correspondences among multiple modalities to understand multi-modal data, which faces some fundamental challenges as listed below:

- 1) **Multi-modal Representation.** The multi-modal representation uses the potential complementary of multi-modality to learn feature representation. The existing efforts either projects the multi-modality into a unified space [38] with VGG [39], ResNet [40], or represents every single modal in its own vector expression space which satisfies certain constraints like linear correlation [41].
- 2) **Multi-modal Translation.** Multi-modal translation learns to translate from a source instance in one modality to a target instance in another. The example-based translation models build bridges between different modality through dictionary [37], [42], while the generative translation models build a more flexible model which can transform one modal to another [43], [44].
- 3) **Multi-modal Alignment.** Multi-modal alignment aims to find the correspondences between different modalities. It can either be directly applied in some multi-modal tasks such as visual grounding, or be taken as a pre-training task in multi-modal pretrained language models [45].
- 4) **Multi-modal Fusion.** Multi-modal fusion refers to the process of joining information from different modalities to perform a prediction [27], where various attention mechanism such as gated cross-modality attention [46], bottom up attention [47] etc. are applied to model the interaction between different kinds of features in the cross-modal module.
- 5) **Multi-modal Co-Learning.** Multi-modal co-learning aims to alleviate the low-resource problems in a certain modality by leveraging the resources of other modalities through the alignment between them [27].

**Multi-Modal Pretrained Language Model.** Based on a large-scale unsupervised multi-modal data set with text-image pairs, much recent efforts works on learning a multi-modal pretrained language model with some designed self-supervised pretraining tasks including masked language model, sentence image alignment, masked region label classification, masked region features regression, masked object prediction, etc [45]. Multi-Modal Pretrained Language

System	MMKG Type	Multi-modalized Knowledge	Source Images	Candidate KGs	Quality Control	Scale
NEIL [19]	N-MMKG	entity, concept, relation	images from search engine	WordNet	semi-supervised classification with labeled seed images	1,152 objects, 1,034 scenes 87 attributes, 1,703 triples (2.5 months)
GAIA [20]	N-MMKG	entity, concept	multimedia news documents	Freebase, GeoNames	object detection, fine-grained classification, heuristic rules	< 457K entities, < 67K triples, < 38K events (including textual and visual ones)
RESIN [21]	N-MMKG	entity, concept, event	multimedia news documents	WikiData	weakly-supervised extraction	< 24 entity, < 46 relations, < 67 events (including textual and visual ones)

(a) Image-based visual knowledge extraction systems that could be used to construct MMKGs by labeling images

MMKG	MMKG Type	Multi-modalized Knowledge	Source KGs	Candidate images	Quality Control	Scale	Images per entity
IMGpedia [22]	N-MMKG	entity, concept, relation	DBpedia	Wikimedia Commons	constructed via DBpedia Commons	12.7M links to KG (with 2.6M entities/concepts), 3000M triples (including 443M triples of 1 visual relation)	>5.6
ImageGraph [23]	A-MMKG	entity, concept	FB15K	search engine	disambiguation by Wikipedia URI	15K entities/concepts	55.8
MMKG [24]	A-MMKG	entity, concept	FB15K, DBpedia, YAGO	search engine	1.entity alignment cross different KGs 2.disambiguation by Wikipedia URI	15K entities/concepts	55.8
Richpedia [25]	N-MMKG	entity, concept, relation	Wikidata	search engine, Wikipedia	1.disambiguation by Wikipedia URI 2.a diversity retrieval model to filter images	29,985 entities/concepts, 172M triples (including 114.5M triples of 3 visual relations)	99.2
VisualSem [26]	N-MMKG	entity, concept	BabelNet	Wikipedia, ImageNet	1.synsets in ImageNet as initial entities pool 2.mining neighbours 3.a image-text matching model to filter noise	89.9K entities/concepts, 13 relations	10.4

(b) MMKGs constructed by symbol grounding

TABLE 2: Mainstream MMKGs (or extraction systems for constructing MMKGs) and their relevant information

Models have proven their effectiveness in improving many downstream tasks [48], [49], [50], [51], [52].

In terms of the Transformer-based fusion process of different modality, the multi-modal pretrained language models can be divided into single-stream models and two-stream models. The single-stream models, such as VLBERT [53] and ViLT [50], input all modal information into a single Transformer encoder for fusion by self-attention modules, and simultaneously learn these representation of different modal data in the same encoder. The two-stream models, such as LXMERT [54], input different modal information into their own encoders and fuse these representation from different modal encoder by an additional cross-attention module. The final output representation not only contains the cross-modal interaction, but also preserves the interaction within each modality.

### 2.3 Discussions

Although there is already much efforts on dealing with various kinds of multi-modal tasks with multi-modal learning techniques and multi-modal pretrained language models, there is still an emerging trend to introduce multi-modal knowledge to help improve the performance of the existing multi-modal tasks. In general, MMKGs could benefit these downstream tasks from the following aspects.

- 1) MMKG provides sufficient background knowledge to enrich the representation of entities and concepts, es-

pecially for the long-tail ones. For instance, auxiliary commonsense knowledge is introduced to enhance the representation of image and text, leading to improved performance of image-text matching [16].

- 2) MMKG enables the understanding of unseen objects in images. Unseen objects pose great challenge for statistic based models. Symbolic knowledge alleviates the difficulty by providing symbolic information about unseen objects or establishing semantic relation between seen objects and unseen objects. For example, external symbolic knowledge is used to guide the generation of captions for unseen novel visual objects with parallel captions [55].
- 3) MMKG enables multi-modal reasoning. For example, an OK-VQA dataset [56], which contains only questions that require external knowledge resources to answer, is built to test the reasoning capability of VQA models.
- 4) MMKG usually provides multi-modal data as additional features to bridge the information gaps in some NLP tasks. Take entity recognition for example, an image could provide sufficient information to identify whether “Rocky” is the name of a dog or a person [57].

To sum up, previous efforts to use multi-modal information is still limited without the support of large scale MMKG. We envision that many tasks can be further improved when a large scale high quality MMKG is available.

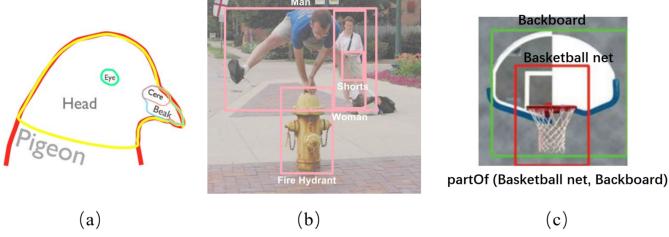


Fig. 2: Examples of labeling images: (a) labeling components after image segmentation in Visipedia [58]; (b) labeling objects with bounding boxes in Visual Genome [59]; (c) labeling two objects where one is a part of the other in NEIL [19].

### 3 CONSTRUCTION

The essence of MMKG construction is to associate symbolic knowledge in a traditional KG, including entities, concepts, and relations etc., to their corresponding images. There are two opposite ways to complete the task: (1) labeling images with symbols in KG and (2) grounding symbols in KG to images. We elaborate the two categories of solutions in Sec. 3.1 and Sec. 3.2 respectively.

#### 3.1 From Images to Symbols: Labeling Images

The CV community has developed many image labeling solutions, which could be leveraged in labeling images with knowledge symbols in KG. Most image labeling solutions learn the mapping from image content to a wide variety of label sets including objects, scenes, entities, attributes, relations, events and other symbols. The learning procedure is supervised by a human annotated data set, which requires the crowd workers to draw bounding boxes and annotate images or regions of images with given labels, as illustrated in Figure 2.

Some well-known image-based visual knowledge extraction systems are constructed as listed in Table 2a, which could be utilized for constructing MMKGs through image labeling. According to the category of symbols to be linked, the process of linking images to symbols could be divided into several fractionized tasks: *visual entity/concept extraction* (Sec. 3.1.1), *visual relation extraction* (Sec. 3.1.2) and *visual event extraction* (Sec. 3.1.3).

##### 3.1.1 Visual Entity/Concept Extraction

Visual entity (or concept) extraction aims to detect and locate target visual objects in images, and then label these objects with entity (or concept) symbols in KG.

**CHALLENGES.** The main challenge with this task lies on how to learn an effective fine-grained extraction model without a large-scale, fine-grained, well-annotated concept and entity image dataset. Although there are rich well-annotated image data in CV, these datasets are almost coarse-grained concept images, which could not meet the requirements of MMKG construction for image annotation data of fine-grained concepts and entities.

**PROGRESSES.** The existing efforts with visual entity/concept extraction could be roughly divided into two categories: 1) object recognition methods, which label a



Fig. 3: The heatmap of word `Soldier` and `Boats` in GAIA [20]. The more relevant between a pixel and a word is , the warmer the color of the pixel is.

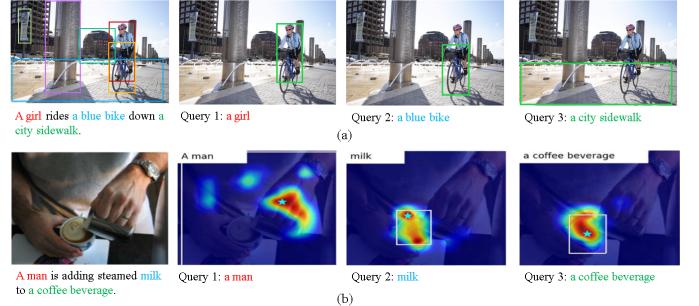


Fig. 4: Two kinds of weakly supervised visual entity extraction: (a) the attention-based method [60]) and (b) the salience-based method [61]. The first method selects the most relevant bounding boxes to given phrases. The second method selects the most sensitive pixels to given phrases.

visual entity/concept by classifying the region of a detected object; and 2) visual grounding methods, which label a visual entity/concept by mapping a word or phrase in a caption to the most relevant region.

1) *Object Recognition Methods.* In earlier work, images provided by users and researchers are usually simple and there is only one object in one image, which can be processed by classification models. But images in our real life may be too complex to be represented with only one label, thus we need to tag different visual units with different labels.

In order to distinguish several visual entities in images, pre-trained detectors and classifiers are needed to label visual entities (as well as attributes and scenes) with their locations in the images. These detectors are trained by supervised data, which comes from public images-text datasets [20] (such as MSCOCO [62], Flickr30k [63], Flickr30k Entities [64] and Open Images [65]) or pre-labeled seed images [19]. During detection, detectors capture a set of region proposals for possible objects and pick out proposals that actually contain objects. At the locations detected by various detectors, such as face detectors based on MTCNN and vehicle detectors based on Faster-RCNN, the pre-trained classifiers recognize candidate visual objects with entity-level (e.g. BMW 320) or concept-level (e.g. Car) labels.

The recognized objects are not directly considered as visual entities due to the large amount of duplication. Many objects are duplicated instances of the same entities at different view points, positions, poses and appearances. Therefore, a visual entity is generated by selecting the most representative visual object. The most common way is to

cluster regions of images, in which the center of each cluster is regarded as a new visual entity [20].

The shortcoming of supervised solutions is obvious that they can only generate a limited number of labels. If the automatic labeling solution is to support large numbers of labels (such as billions of entities), it requires a lot of pre-processing work, such as predefined rules, predetermined lists of recognizable entities, pretrained fine-grained detectors and classifiers, etc., which deteriorates the scalability of the solution [20].

2) *Visual Grounding Methods.* In visual entity extraction, training detectors needs a large amount of labeled data with bounding boxes and pre-defined schema with a fixed set of concept [66], which is difficult to be used for large-scale visual knowledge acquisition. Fortunately, there are a large number of image-captions pairs from the Web (e.g. news websites) to weakly supervise the extraction of visual knowledge without relying on the labeled bounding boxes. Therefore, the visual entity extraction problem is reduced to an open-domain visual grounding problem, which aims to locate the corresponding image region of each phrase in a caption so as to obtain visual objects in an image with their labels.

When extracting information from weakly supervised image-caption pairs, we often directly select active pixels for a given word as the region of visual objects based on the spatial heatmap (e.g. the heatmap in Figure 3). With text and image representations shared in the same semantic space, the heatmap of each phrase can be learned by attention-based methods and saliency-based methods as cross-modal weight, as shown in Figure 4. At training time, saliency-based methods directly treat the sensitivity of pixels to a given phrase as the value of heatmap by gradient computation. Attention-based methods treat the cross-modal relevance as the value of heatmap, which is more popular compared to saliency-based methods. For example, the heatmap in GAIA [20] is generated based on the similarity between image regions with each entity mention in a caption, and that in [67] is generated based on the similarity between image regions and possible event argument role types. At test time, the heatmap is thresholded to obtain a suitable bounding box of a visual object. If there is no overlap between the bounding boxes of the existing visual entities/concepts in KGs and the new bounding box, the bounding box will be created as a new visual entity or concept.

Although the open domain oriented visual grounding doesn't rely on labeled data with bounding boxes, in practice human verification is still needed due to mismatching. Some efforts attempts to add constraints on common concepts, relations and event arguments into training stage to increase supervision information. The precision of visual grounding is less than 70% in the work related to the construction of MMKG [20]. The located visual objects via visual grounding could be entity (e.g. Barack Hussein Obama), concepts (e.g. place, car, stone), attributes (e.g. red, short). However, the inconsistency of semantic scales of images and text may lead to incorrect matching. For example, troops may be mapped to *several individuals wearing military uniforms*, and Ukraine (country) may be mapped to *a Ukrainian flag*, both of which are just relevant



Fig. 5: Weakly supervised visual entity extraction via multi-modal pre-trained language models. This figure shows the most relevant regions of an image to given words in a caption through self-attention mechanism of ViLT [50].

but not equivalent.

**OPPORTUNITIES.** As the emerging of multi-modal pre-trained language model as covered in Sec. 2.3, the powerful capability of representation brings more opportunities to extract common entities and concepts. The mapping of image patches and words can be directly visualized in the self-attention maps of the model without additional training. An example of the prediction of ViLT [50] is shown in Fig. 5. It is proved that multi-modal pre-trained language models, such as CLIP [68], trained on hundreds of millions of web-scale image-text data are highly accurate on famous figures and landmark buildings [69], which will reduce much workload on data collection and model training when constructing a MMKG of figures or buildings. Some pre-trained vision transformer models already have strong capability of visual object segmentation and focus on the foreground object even in highly ambiguous situations, such as DINO [70], which will improve the performance of locating visual objects and aligning cross-modal knowledge.

### 3.1.2 Visual Relation Extraction

Visual relation extraction aims at identifying semantic relations among detected visual entities (or concepts) in images, and then labeling them with the relations in KGs [19].

**CHALLENGES.** Although visual relation detection has been studied extensively in CV community, most of the detected relations are superficial visual relationship between visual objects such as (*person, standing on, beach*). Differently, for the purpose of constructing MMKG, the visual relation extraction task aims to identify more general types of semantic relations that are defined in KGs such as (*Jack, spouse, Rose*).

**PROGRESSES.** The existing efforts on visual relation extraction can be roughly put into two categories including rule-based relation extraction and statistic-based relation extraction. Some other work mainly focuses on long-tail relation and fine-grained relation, which will also be covered in the following.

1) *Rule-based Relation Extraction.* Traditional rule-based methods mainly focus on some specific types of relation, such as spatial relation [71], [72] and action relation [73], [74], [75], [76], [77], [78]. The criteria are usually predefined by experts and the discriminative features are scored and selected by heuristic methods.

In rule-based methods, the relations to be detected are defined in terms of the types of labels and the relative locations of regions. For examples, if the bounding box of one object is always within that of another object, there may be a *PartOf* relation between them. Table 3 lists several visual relation detected in NEIL [19]. During the extraction

in NEIL, the detected relation between a pair of objects are in turn an additional constraint for new instance labeling. For example, “*Wheel is a part of Car*” indicates that it is more possible for a *Wheel* appears in the bounding box of a *Car*. Rule-based methods provide highly accurate visual relations, but they rely on a large amount of manual work. So it is not practical in large-scale MMKG construction.

2) *Statistic-based General Relation Extraction*. The statistic-based methods encode features such as visual features, spatial features, and statistics of the detected objects into distributed vectors and predict the relation between the given objects by a classification model. Compared to rule-based methods, statistic-based methods are able to detect all relations that have appeared in the training set.

Some work has proved that predicting the predicates heavily rely on the categories of subjects and objects, but subjects and objects are not dependent on predicates and there is also no dependency between subjects and objects [79]. For example, in triple (*person*, *ride*, *elephant*), *person* and *elephant* indicate that the relation might be *ride* rather than *wear*. Thus to utilise the dependency, [79], [80], [81] add language priors of language models into statistic model by objects’ labels and [82] set a stricter constrain that the hidden layer representation of a triple should satisfy *subject + predicate ≈ object*. It is embarrassing that the language model brings a lot of improvement but the visual information contributes very little [79].

Detected objects and relations in an image could be represented as a graph. The graph structure enables the edges to get more messages from other nodes and edges to classify the relation with higher accuracy. For example, [83] represents objects and relation as two complementary sub-graphs, where nodes are iteratively updated according to the values of the surrounding edges and vice versa. [84] used attentional GCN to learn the context objects and edges.

3) *Long-tail and Fine-grained Relation Extraction*. Although statistics-based methods are able to detect general relations, it is difficult to detect long-tail relations. The reason is that the biased datasets make it more possible to predict the relations with a large number of samples. In order to eliminate the effect of unbalanced samples in the training sets, [85] proposed a new unbiased metric (Mean Recall@K) to average the recall of all types of relations instead of all samples, and avoiding the neglect of relations with only few samples. There are much other work focusing on detecting relations with few samples by transfer learning [86], few-shot learning [87] and contrastive learning [88], still being limited to the feature fusion of hidden layers.

Fine-grained relation is a kind of long-tail relation. Existing studies on long-tail relation problem from the perspective of feature fusion fails to distinguish fine-grained relations well. For example, models tend to predict *on* instead of fine-grained relation *sit on/walk on/lay on*. To get more fine-grained and informative unbiased predictions, [89] uses counterfactual causation instead of conventional likelihood to remove the effect of context bias.

It is more difficult to detect more complex and fine-grained relation such as human-object interaction and action detection. Because the pose of a person is determined by many components of the body. For example, there is slight difference between images of (*person*, *play*, *violin*)

and (*person*, *hold*, *violin*). In the first example, the *violin* is usually placed on a person’s neck, with the left hand pressing the strings and the right hand holding the bow, and in the second example the *violin* may be absolutely different poses. In early studies, action is defined as a serial of poses of different parts of the body, and the discriminative features are mined by heuristics approaches [73], [74], [75]. In the current the statistics-based detection, the discriminative feature are filtered by stricter contrastive loss functions [90], which is obviously still too coarse.

**OPPORTUNITIES.** Despite much existing work, there still leaves many challenging issues unsolved. For instance: 1) *Visual Knowledge Relation Judgement*. Many visual triplets extracted from images only describe the scene of the image, which are unqualified to be taken as visual knowledge since they are not widely accepted facts. The challenges (also opportunities) lie on how we recognize the triples of visual knowledge from the triples of scene information. 2) *Relation Detection based on Reasoning*. Existing relation detection methods predict the relations by a hidden unified representation fusing visual features and language priors. We can’t explicitly describe the basis of prediction. [91] builds a human action datasets to help predict an action by body part states. For example, if there is a *person* and a *football* in an image and (*head*, *look at*, *sth*) (*arm*, *swing*, *-*) (*foot*, *kick*, *sth*) are meanwhile satisfied, the action will be judged as (*person*, *kick*, *football*). Unfortunately, this dataset is build manually. We need to automatically summarize the chain of reasoning for relation detection.

### 3.1.3 Visual Event Extraction

An event is usually defined as the dynamic interaction among arguments [92], including a trigger and several arguments with their corresponding argument roles. A trigger is a verb or a noun indicating the occurrence of an event, an argument role refers to the semantic relation between an event and an argument such as *Time*, *Person* and *Place*, and the arguments are entity mentions, concepts or attribute values. The traditional event extraction task aims to predict the event types by triggers and then extract their arguments according to the pre-defined schema of the event. Similarly, the visual event extraction can be also divided into two sub-tasks: 1) to predict the visual event types; and 2) to locate and extract objects in source images or videos as visual arguments [21], [67], [92], [93].

**CHALLENGES.** There are several challenges with the task: 1) Visual event extraction requires pre-defined schema for different event types, but a large number of visual events have not been defined by experts. How to mine visual pattern as event schema automatically? 2) How to extract the visual arguments of a visual event from images or videos?

**PROGRESSES.** The existing work on visual event extraction mainly focuses on two aspects: 1) visual event schema mining, which detects and labels the most relevant visual entities (or concepts) as a new schema; 2) visual event arguments extraction, which extracts argument role regions from visual data according to the event schema.

relation type	example	images	relation type	example	images
Concept-Concept	Keyboard is a part of Laptop.		Scene-Entity	Ferris wheel is found in Amusement park.	
Entity-Concept	BMW 320 is a kind of Car.		Scene-Attribute	Alleys are Narrow.	

TABLE 3: Examples of visual relations detected in NEIL [19]

1) *Visual Event Schema Mining*. In CV community, there is a similar task called situation recognition task, the schema of which includes the main activity (event type), the participants (arguments), the roles theses participants play in the activity (argument role). Some datasets, such as SituNet [94] and SWiG [95], have defined a lot of visual event schemas. For example, the event Clipping have argument roles like Agent, Source, Tool, Item, Place and in an image of *clipping a sheep’s wool* they are respectively Man, Sheep, Shears, Wool, Field. The task mainly aims to recognize an visual event rather than locating and extracting its visual arguments. The visual arguments detection relies on the co-occurrence [94], [95] and relative positions [96] of detected objects.

However, in large-scale visual event extraction, such as news, the visual schemas of many events have not yet been manually defined, which requires a lot of experts’ work. A large number of image-caption pairs from Internet make it possible to mine and label the visual pattern for event schemas. Thus this task is reduced to find a frequent itemset of visual patterns which indicate the correct event type from the images of a given event. The collection of images of an event can be retrieved from the image-caption pairs with the triggers of the event as queries. Through visual grounding methods, the candidate image patches are labeled by words or phrases in captions. Heuristic approaches (e.g. Apriori algorithm) can be utilized to mine frequent visual image patches to find association rules for predicting the event type by visual patterns [92], [97].

The mining and labeling methods are able to correct wrong arguments or add missing arguments in manual defined visual events schemas. For example, an ontology expert may consider Explosion and Weapon as important items in the schema of event Attack, but in some news corpus, these concepts are not discovered and Smoke and Police appears much more frequently, which is not expected in advance [97].

2) *Visual Event Arguments Extraction*. The visual event arguments extraction is actually a task of extracting a group of visual objects with the constrain of relations. Visual arguments are able to be labeled by fully supervised methods like object recognition or weakly supervised methods like visual grounding. According the two sub-tasks of visual event extraction, the event types are classified by the global features of events’ images and the event arguments are extracted as the most sensitive local region to the event type.

However, during weakly supervised methods, we are not sure whether the relations among extracted visual objects are consistent with the relations in text. Thus, the relations in both visual arguments and text arguments

should also be aligned respectively. [67] aligns the situation graph [94] extracted from an image of an event to the abstract meaning representation graph (AMR graph) [98] representing the semantic structure of the caption of this event in terms of the semantic and categories of cross modal arguments. Many constrains on semantic, event type, event argument role and the consistency of visual and text information are also added into joint extraction [21], [67].

Compared to images, videos are more suitable for event extraction, because the temporal bounding box of an event may across the video and all arguments may not be shown in a single frame. To simplify the task, [93] extracted arguments from three key frames derived from short video segments including only one event, and the key frames are the most matching ones to the captions of the videos.

**OPPORTUNITIES.** The research on this task is still in an early stage, and there are still many problems worth exploring. For instance: 1) The extraction of sequential events from a long video containing multiple events has not yet been addressed. 2) *Video Event Extraction with multiple Sub-events*. For example, the event Making Coffee is divided into a sequence of steps, such as Cleaning coffee machine → Pour in the coffee beans → Turn on the coffee machine and each step can be also considered as an event. The sequential steps need to be extracted and listed by the timeline of the steps, which are difficult to be solved by current methods.

### 3.2 From Symbols to Images: Symbol Grounding

Symbol grounding refers to the process of finding proper multi-modal data items such as images to denote a symbol knowledge exists in a traditional KG. Compared to the image labeling way, the symbol grounding way is more widely applied for MMKG construction. Most of the existing MMKGs are constructed in this way, as listed in Table 2b.

In the rest of this subsection, we cover the process of grounding symbols to images in several fractionized tasks: *Entity Grounding* (Sec. 3.2.1), *Concept Grounding* (Sec. 3.2.2) and *Relation Grounding* (Sec. 3.2.3).

#### 3.2.1 Entity Grounding

Entity grounding aims to ground entities in KGs to their corresponding multi-modal data such as images, videos and audios [12]. The existing work mainly focuses on grounding entities to their corresponding images.

**CHALLENGES.** The main challenges of grounding entities to images are the following: 1) How to find enough images with high quality for entities at a low cost? 2) How to select the images that best match an entity from much noises?

**PROGRESSES.** There are two major sources to find images for entities: (1) from *online encyclopedia* (such as Wikipedia) and (2) from the Internet through *Web search engines*.

1) *From Online Encyclopedia.* In Wikipedia, an article is usually describing an entity with images and other multimodality data. Wikipedia and DBpedia provide a lot of facilities (such as Wikimedia Commons<sup>1</sup>) to help build the connection between an entity in DBpedia and corresponding images or other modality data in Wikipedia. It is easy for researchers to use online encyclopedia like Wikipedia to build a first version of a large scale MMKG.

However, the encyclopedia-based approach unfortunately has three major disadvantages: 1) First, the numbers of images for each entity is limited. The average number of images per entity in Wikipedia is 1.16 [22]. 2) Second, many images of entities in Wikipedia are only relevant to its corresponding entity, but not exactly referred to the entity. For example, there are several images of *animals*, *buildings*, *plaques*, *carvings* in images of Beijing Zoo in Wikipedia, which could easily lead to semantic drift. In addition, the images of non-visual entity bring mistakes. For example, in the Wikipedia article of *Gaussian Progress*, there is an image of *Gaussian processes with different prior conditions*, which actually should not be mapped to any image. 3) Third, The coverage of MMKG built from Wikipedia still needs to be improved. English Wikipedia has 6 millions of entities (articles), which is the upper bound of the capacity of the MMKG harvested from English Wikipedia. According to our investigation, nearly 80% Wikipedia articles in English have no corresponding images, and only 8.6% of them have more than 2 images.

2) *From Search Engines.* To improve the coverage of a MMKG, search engine based solutions are proposed. We can easily find images from the search result of commercial search engine by specifying entity names as queries. The top ranked result image in general has a large possibility to be the correct image of the entity to be searched. Thus we can select these images for the entity to be searched. Compared to the Wikipedia based approach, the coverage of MMKG is significantly improved in search engine based approach.

However, search engine based approach is easy to introduce wrong fact into MMKG. It is well recognized that the search engine results might be noisy. An additional reason is that it is not trivial to specify the search key words. For example, the search query “Bank” is not good enough to find the image for Commercial Bank, since it also incurs the images of River Bank. Hence, a lot of efforts has been made to clean candidate images. The query words are usually extended for disambiguation by adding parent synsets [99] or entity types [24]. The diversity is also non-negligible issue when selecting the best images for the entity. An image diversity retrieval model is trained to remove redundant similar images so that the grounded images are as diverse as possible [25].

Due to decoupling of entities and their visual features during construction, the MMKG based on entity grounding has the ability to distinguish visually similar entities, as shown in Fig. 6. Entity grounding methods make it possi-

1. a multi-media dataset linking to Wikipedia articles, <https://wikimediawiki.org/our-work/commons/>



Fig. 6: (a) Similar visual entities: Natalie Portman and Keira Knightley; (b) Similar visual concepts: fire fighter and trash collector.

ble to build a domain-oriented fine-grained MMKG (e.g. a movie/product/military MMKG).

Compared to the encyclopedias-based approach, search engine based approach is better in coverage but worse in quality. The two approaches are often used together since in most cases the knowledge acquired by these two approaches complement with each other [25], [100]. For example, the coverage of MMKG harvested from Wikipedia can be improved by collecting more images for each entity from search engines [25] or mapping each image to all the entities it contains to expand the number of entities’ images [100].

**OPPORTUNITIES.** There are many unsolved problems in this direction. 1) Entities are grounded into several images, each of which is only an aspect of the entity. For example, the images collection of a person may be images of different ages, life photos, event photos, single photos and family photos. How to determine the most typical subset? 2) Real world entities are multi-faceted, and it is desirable to associate an entity with multiple images under different contexts. This motivates us to propose a new task *multiple grounding* which selects the most related images from the entity given a specific context. This problem could be formulated as  $\underset{i}{\operatorname{argmax}} P(i|c)$ , where  $e$  denotes an entity and  $c$  denotes a specific context.  $I_e$  denotes an image pool of entity  $e$  and  $P(i|c)$  is the probability that image  $i$  is an appropriate image for entity  $e$  conditioned on the given context  $c$ . For example Donald Trump, the 45th and current president of the United States, has a lot of different images that can be collected from web. But as shown in Figure 7, any single image is not appropriate for all the different contexts. Thus, Trump should be multi-grounded when we construct the knowledge graph. However, mapping different aspects of entities to the most related images in different contexts is not a trivial task. First, the image pool of an entity is difficult to build, because the completeness of the image pool cannot be guaranteed and it is easy to miss some related images for certain contexts. Second, disambiguating images for the entity for a specific context is challenging, because the context is usually noisy and contains sparse information, and more background information is needed to guide the acquisition of semantic information. Finally, as a new task, the lack of labeled data is a big issue .

### 3.2.2 Concept Grounding

Concept grounding aims to find representative, discriminative and diverse images for visual concepts.

**CHALLENGES** Although some visually unified concepts (such as man, woman, truck and dog) can also be grounded to images with the entity grounding methods introduced in

- S1: In 1964, *Trump* enrolled at Fordham University.  
 S2: In 1971, *Trump* was named president of the family company and renamed it The Trump Organization.  
 S3: *Trump* registered as a Republican in Manhattan in 1987.  
 S4: *Trump* is the wealthiest president in U.S. history, even after adjusting for inflation

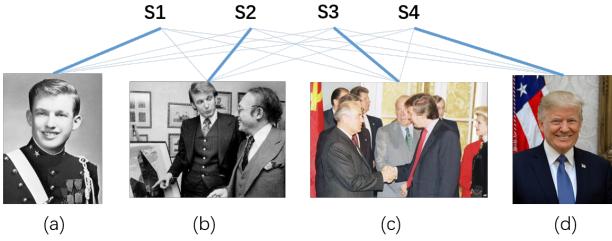


Fig. 7: Sentences and Images about Trump. (a) Photos of Trump's early life, (b) Trump as a businessman, (c) Trump as a politician, and (d) Trump as the president of USA. Obviously, (a) is more related to S1, (b) is more related to S2, (c) is more related to S3 and (d) is more related to S4.

Sec. 3.2.1, the symbol grounding to the other concepts faces new challenges: 1) Not all the concepts could be visualized properly. For example, *irreligionist* can't be grounded to a certain image. How to distinguish visualizable concepts from non-visualizable ones? 2) How to find representative images for a visualizable concept from a group of relevant images? Note that the images of a visualizable concept might be very diverse. For examples, when it comes to *Princess*, people often think of several diverse images, which are Disney princesses, ancient princesses in historical movies or modern princesses in the news. Therefore, we have to take the diversity of images into consideration.

**PROGRESSES.** In response to the above challenges, related researches are divided into three tasks: visualization concept judgment, representative image selection and image diversification.

1) *Visualization Concept Judgment*. The task aims to automatically judge visualizable concepts and is a new task to be solved. [101] discovers that only 12.8% of the synsets of *Person* subtree have well-accepted imageability (score is greater or equal to 4 and total score is 5) and many of rest synsets have no corresponding visual descriptions. For example, *Rock star* is imageable, and *Job candidate* is non-imageable. So what are the criteria for recognizing visual concepts? The manual annotation used in [101] is obvious unpractical in the construction of a large-scale MMKG.

In order to automatically judge visual concepts, there has been much efforts based on syntax and semantics. [102] thinks that abstract nouns concepts are non-visual so that TinyImage dataset [102] removes all hyponyms in the subtree of *Abstraction* in WordNet and only collects images for non-abstract noun concepts. However, these methods are not very accurate. For example, *Anger* or *Happiness* can be grounded into an image of a person who feels angry or happy. Since the images come from the Internet, it is possible to use search engine hits to recognize visual concepts. For example, a word might be visible if the number of Google image hits is larger than the number of Google web hits [103].

In addition, some characterises of high-qualified images of visual concepts could be used to recognize visual con-

cepts, such as representativeness and discriminativeness. [104] believes the foreground of representative images are similar and are easily to be separated from background and have a small inter-class variance. Thus in turn [104] trains a classifier to select the concepts whose images collection have these characterises.

2) *Representative Image Selection*. By the methods of Sec. 3.2.1, we get a collection of images for each visual concept. In this section, we focus on selecting visually representative and discriminative images in the collection.

The task essentially aims to re-rank the images according to their representativeness. The representativeness of images are scored in terms of results of cluster-based methods, such as K-means, spectral clustering, etc. The smaller the variance within a cluster, the higher the scores of images in the cluster. After reranking the scores of representativeness of images, the top may be representative images. In addition, the expected images are also constrained by rules to distinguish different clusters. For example, [105] adds a new metric to rank images together with similarity within clusters, which is the ratio of inter-class distances and intra-class distances, and the bigger a ratio, the more discriminative the image is.

The captions and tags of images from search engines could also be utilized to evaluate the representativeness and discriminativeness of images in the level of semantic. Captions and tags provide semantic information that images do not have. For example, a photo of *Icelandic landscapes* and a photo of *British landscapes* may look similar, but text tags can help us distinguish their difference of concepts. In [103], [106], [107] tags are clustered based on semantic features and images are re-assigned into each cluster according to their tags' semantic clusters.

3) *Image Diversification*. The task requires the images which concepts are grounded to should balance diverse and relevant. The images should also be re-ranked after clustering, but the difference from representative image selection is that, we want to show the results of as many clusters as possible. Specifically, in each selection, try to select images from cluster which have not been selected.

There are two types of scores to rank the priority of selection: diversity scores and relevance scores, where diversity scores evaluate the topics of images and relevance scores penalize the difference of images to avoid semantic drift. To combine the two conflicting scores, [108], [109] use Max-Min methods to choose candidates: assign higher score to images that are not similar to the selected set, and choose the dissimilar one with the highest score among the remaining similar ones.

We can also resolve the ranking problem by graph algorithms. A set of images could be represented as a graph, where images are nodes and visual similarities between images are weights of edges. Thus, the ranking of representative images reduces to finding an optimal path in a fully connected graph with the respect to re-weighted values of edges. [110] uses dynamic programming to search for the optimal sequence in an image graph, where the value of edges are a joint criterion combining diversity score and relevance score. Markov random walk is also used for the optimal sequence in [103], [111], where [111] weight the values by Max-Min methods and [103] reassigned the visits

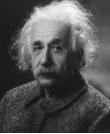
concept type	visualizable concept	non-visualizable concept
example	Surgeon	Physicist
image		

TABLE 4: Examples of visualizable concept grounding and non-visualizable concept grounding via entity grounding. The non-visualizable concept Physicist can be grounded to the photo of Einstein since Einstein is a typical entity of Physicist.

value between nodes according their source clusters by a two-layer graph model.

These studies concentrate on the field of text-image retrieval and few of them are related to MMKGs. There are still many unsolved bias in the diversity of images of concepts derived from the Internet on gender, race, color and age, and the problem now relies heavily on crowdsourcing [101].

**OPPORTUNITIES.** As a fledgling area, many unsolved problems are leaving for future research. We give two examples below:

1) *Abstract Concept Grounding.* Previous work on concept visualization judgement seldom considers abstract concepts. But the abstract concepts could also be grounded with images. For example, Happiness are usually associated with *smile*, and Anger are usually associated with *an angry face*. Some abstract nouns have diverse but fixed visual association, such nature, human and action. For example, in [112] the images of Beauty are associated with following word clusters: *woman/girl, water/beach/ocean, flower/rose, sky/cloud/sunset*. Similarly, the image of Love are associated with following word clusters: *baby/cute/newborn, dog/pet, heart/red/valentine, beach/sea/couple, sky/cloud/sunset, flower/rose*. It can be seen that some abstract nouns often have generic and fixed images in terms of sentiment and have discriminative images in terms of semantic.

2) *Gerunds Concept Grounding.* Gerunds are a special kind of nouns, which could be transformed into verbs, such as *singing → sing*. [76] grounds many gerunds to images through crowdsourcing, such as *arguing with, wrestling with and dancing with*. These verbs about human interaction are sensitive to the features of person's body angle, gaze angle, position of the joints and expression.

3) *Non-visualizable Concept Grounding via Entity Grounding.* If a concept is non-visualizable but the entities of the concept could be visualized, the concept could also be grounded via its entities. For instance, a reasonable selection of the grounded image to such a concept is to use the image of the concept's most typical entity. As shown in Table 4, we use a photo of Einstein to ground the concept Physicist. It is reasonable since most of us will think up with Einstein when we mention a Physicist. However, there are still a lot of unresolved questions: (a) In general, different person think up with different typical entity for a concept, so we should address such subjectivity in concept grounding. Whether an entity is a typical one in the constrain of its concept? (b) We should choose several typical

entities' images to present that concept. How to summarize and select typical entities to represent concepts? (c) Whether should we abstract common visual features from multiple images of entities?

### 3.2.3 Relation Grounding

Relation grounding is to find images from an image data corpus or the Internet that could represent a particular relation. The input could be one or more triples of this relation, and the output is expected to be the most representative images for the relation.

**CHALLENGES.** When we take a triple as a query to retrieve images for the relation, the top-ranked images are often more relevant to the subject and object of the triple, but not that relevant to the relation itself. How to find images that could reflect the semantic relation of the input triples?

**PROGRESSES.** Existing studies on relation grounding focus on spatial or action relations such as *left of, on, ride and eat*.

While textual queries can be represented as structured data in the format of *(subject, relation, object)* through AMR graph [98], candidate images could also be structured into a scene graph [113]. Then, the structured text and structured images could be matched in fine-grained level by means of text-image matching or graph matching, which are introduced in the following.

1) *Text-Image Matching.* In text-image matching tasks, text and images are usually represented as vectors in a unified semantic embedding space. The images that best match the query is found by the similarity score of cross-modal representations. The cross-modal representations are usually fused by the mechanism of attention, so the disadvantage of global representation is lacking semantics of explicit fine-grained relations [35]. In addition to representation-based retrieval, a more convenient method is caption-based retrieval like search engines on the Internet. The disadvantage of caption-based retrieval is that the visual features have not been used for matching.

To represent explicit relation between objects, many studies concentrate on a image encoder considering local structure of images. The final image representations are the fusion of global visual features, local structure features and text aligned embeddings [78], [114], [115]. In [78], all one-order (entity or concept), two-order (attribute or action), three order (triple) facts are modeled by a unified setting  $(s, p, o)$ , which are respectively represented by the output of different branches of a multiple layer image encoder. [115] uses a scene graph to represent all triple  $(s, p, o)$  in an image and uses graph convolutional neural network to learn the visual relations. Finally, all the visual representations with relation features learned for each image have to be close to the text embeddings of corresponding words in the captions. Thus, the matching images can be directly retrieved by using a triple as a query instead of a sentence.

Multi-modal pre-trained language models are a new alternative to image encoders that consider objects (entities or concepts) and triples. For each image-caption pair, a scene graph parser is employed to generate a scene graph with objects, attributes and relations from the caption of an image, then UNIMO [51] randomly replaces the object, attribute

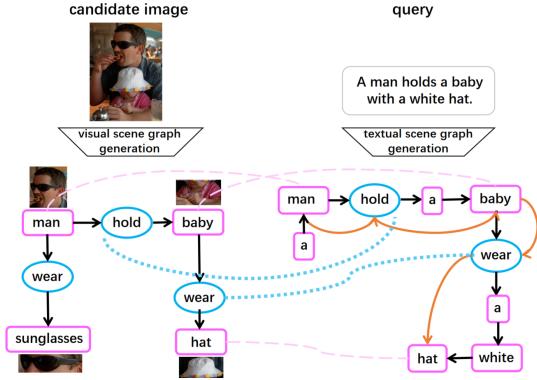


Fig. 8: Relation grounding by graph matching in [116].

and relation nodes of a scene graph with a different object, attribute or relation from the corresponding vocabularies to generate a large number of hard negative samples. ERNIE-ViL [52] enhances the capability of visual and language model by adding three pre-training tasks, object prediction, attribute prediction and relation prediction.

*2) Graph Matching.* We expect the relation grounding through explicit matching of objects and relations, rather than implicit matching of unified cross-modal embeddings. A more convenient method is caption-based retrieval like search engines on the Internet, matching tokens of entities and relations between the query and captions. The disadvantage of caption-based retrieval is that the visual features haven't been used for matching. For example, Richpedida [25] proposes a very strong assumption that if there is a pre-defined relations (e.g. `nearBy` and `contain`) between two entities in the Wikipedia descriptions, the same relations also exist between two entities' corresponding visual entities. But in reality, it is more likely that the two objects are not simultaneously appear in one image. Even if they do, there may be no expected relation in one image.

If we represent the textual query and candidate images into graphs, the relation grounding task turns into a task of graph matching, as illustrated in Figure 8. An image could be structured into a graph in which nodes are object and edges are relations. The dependencies in the textual query could be modeled as the dependency parse tree, which is also a graph. A simple solution is that to only match objects and co-occurring relations in two graphs without predicting the relation types [113]. [113] assumes that if there are a relation between two entities, the relation is considered to be a match, which is also a strong assumption. Obviously, relation prediction module is essential. [116] respectively represents two scene graph by GCN, in which objects update from themselves and relation nodes update from the aggregations of its neighbors. When predicting the similarity of two graphs in different forms is measured respectively: objects nodes matching and relations nodes matching.

**OPPORTUNITIES.** Existing studies mainly focus on grounding spatial relations and action relations, which could be observed visually in images. However, most of the other relations may not be visually obvious in images such as `isA`, `Occupation`, `Team` and `Spouse`. There usually lacks training data for these relations, thus it is difficult to

train models to retrieve images with the above two kinds of solutions.

## 4 APPLICATION

After a systematic review on MMKG construction, this section explores how the knowledge in MMKGs can be applied to and benefit a wide variety of downstream tasks. For a quick overview, Table 5 lists some mainstream application tasks, their benchmark datasets, and the advantages brought by MMKGs. We categorize such tasks into (i) in-KG applications (Sec. 4.1) and (ii) out-of-KG applications (Sec. 4.2), discussed as follows.

### 4.1 In-MMKG Applications

In-MMKG applications refer to the tasks that conducted within the scope of the MMKG where the embeddings of entities, concepts and relations are already learned. Thus, before introducing in-MMKG applications, we briefly go through the distributed representation learning of the knowledge in MMKGs, also named as MMKG embedding.

Basically, the MMKG embedding models are developed from the embedding models on conventional KGs, i.e., *distance-based* models [133], which consider that the head entity and tail entity of the same triplet should be close in the projection space, and *translation-based* models, TransE [134] and its variants [135], [136], [137], which should conform to the assumption:  $t \approx h + r$ .  $h$ ,  $t$ ,  $r$  is respectively the vector representation of head entity, tail entity and relation in a triplet. There are two additional issues in dealing with the data of multi-modalities: how we effectively encode the vision knowledge and information contained in images, and how we fuse knowledge of different modalities. 1) *Vision Encoders.* Although there are many ready techniques for image information encoding in CV, with the development of deep learning, hidden features from Convolutional Neural Networks are the main image embeddings used in MMKG representation [138], [139], [140], while other explicit visual features such as Gray Histogram Descriptor (GHD), Histogram of Oriented Gradients Descriptor (HOG), Color Layout Descriptor (CLD) can hardly be leveraged in MMKG representation. 2) *Knowledge Fusion.* To fuse the knowledge embeddings of multi-modalities, various fusion ways are considered, including simple concatenation, average of multiple modalities' embeddings and normalization-based or weighted SVD and PCA [139]. While some methods [139] take the fused results as the MMKG embedding directly, the other methods [140] further train the uni-modal representations on a well designed objective function.

In the following, we introduce four well-studied in-MMKG applications including link prediction, triple classification, entity classification, and entity alignment.

#### 4.1.1 Link Prediction

Link prediction in MMKG [134], [136], [141], [142] aims to complete a triple  $(h, r, t)$  when one of the entities in  $h, r, t$  is missing, i.e., predicting  $h$  in  $(?, r, t)$  or predicting  $t$  in  $(h, r, ?)$ . A similar task is to predict the missing relation between two given entities, i.e., predicting  $r$  in  $(h, ?, t)$  [138], [143], [144].

Multimodal Application	Benchmark Datasets	Advantages with MMKG
Entity Recognition and Linking	Twitter2015 [57] Twitter2017 [117] Weibo [118]	1.background knowledge provides deep features of images 2.images provide necessary complementary information, help to capture the relationship among mentions and entities 3.learn distributed representations for each entity with multi-modal data
VQA	GQA [119] Visual7w [120] OK-VQA [56] FVQA [33] KVQA [121] KB-VQA [122]	1.provide knowledge about the named entities and their relations in the image, leading to deeper visual content understanding 2.conduct the reasoning process and predict the final answers in a more explicit way with symbolic knowledge from MMKG 3.refine the answers with more interpretability and generality
Image-text Matching	Flickr30k [63] MSCOCO [62] Visual Genome [59]	1.expand more semantic concepts 2.introduce informative relationships between visual concepts by constructing scene graphs and enhance image representations 3.enhance the reasoning and inference capabilities of multi-modal data with graph-structured information
Image Tagging	NUS-WIDE [123] MIRflickr [124]	help disambiguation the concept and relate them better to images
Image Captioning	MSVD [125] MSCOCO [62] NYTimes800k [126] GoodNews [127]	1.enable the understanding of unseen objects with MMKG symbolic knowledge 2.leverage MMKG for relational reasoning to generate more accurate and reasonable captions 3.capture fine-grained relationships between entities in different modalities
Visual Storytelling	VIST Dataset [128]	1.triples in MMKG provide explanation and traceability for described facts 2.provide a strong logical inference between images for more fluent story
Recommender System	MovieLens [129] IntentBooks [130] Dianping [131] KKBOX [132]	1.provide background knowledge for items with rich semantics and help solve the CF cold-start problem 2.learn through rich path semantics across different modalities in MMKG and produce an interpretable and explicit recommendation 3.construct personalized MMKG for items and model entity relation reasoning between them

TABLE 5: Benchmark datasets for their corresponding multimodal applications incorporating MMKGs.

Conventionally, Link prediction on MMKG can be processed with a simple ranking procedure, which finds the best fit entities to complete a triple from all the candidate entities. Compared to the task in traditional KGs, the images attached to entities and relations in MMKGs could provide extra visual information to enhance the embedding learning quality. For instance, the images of a person might provide evidence for the person’s age, profession, and designation [140].

Some other work IMAGEgraph [23] proposes to express the relation prediction between unseen images and multi-relational image retrieval as visual-relational queries, such that these queries could be leveraged for MMKG completion. Compared to the conventional way, IMAGEgraph performs better on the relation and head/tail entity prediction tasks and is able to be generalized to unseen images, to answer some zero-shot visual-relational queries. For example, given an image of an entirely new entity that is not part of the KG, this approach can determine its relation with another given image which we do not know the underlying KG entity.

Similarly, MMKG [24] constructs three datasets to predict the multi-relational links between entities, with all the entities associated with numerical and visual data. However, it only focuses on the sameAs link prediction task and answer such queries for MMKG completion. Three quite heterogeneous knowledge makes MMKG an important benchmark to measure the performance of multi-relational link prediction methods and validates the hypothesis that different modalities are complementary for the sameAs link prediction task.

#### 4.1.2 Triple Classification

Triple classification aims to distinguish correct triples from incorrect ones, which can also be seen as a sort of KG completion task. Based on the embedding model learned on a MMKG, each triple could be calculated with an energy

score  $E(h, r, t)$ . Different threshold  $\delta_r$  is set for each relation  $r$ , and a triple will be predicted to be negative if its energy score is higher than  $\delta_r$ . To prepare the training data for the classification model, correct triplets are corrupted by replacing one of the  $h, r, t$  to generate negative data [138], [143], [144].

#### 4.1.3 Entity Classification

Entity classification categorizes entities into different semantic categories, i.e., concepts of different grains in the MMKG. Entity classification can also be regarded as a special link prediction task, where the relation is set to `IsA` and the tail of the triple waiting to be predicted is a concept in the MMKG.

Various entity classification models have been proposed for traditional KGs, which could also be adopted in MMKGs. But the rich multi-modal data for entities and concepts in MMKGs cannot be fully utilized without a good MMKG embedding model. For instance, some efforts [145] works on learning embeddings for entities and concepts from several different types of modalities and then encode them to a joint representation space.

#### 4.1.4 Entity Alignment

Entity alignment works on aligning entities that refer to the same real-world identity in different MMKGs. It is a viable way to integrate two MMKGs into one when there are overlaps between them.

The core idea is to learn representations for entities in different KGs, and then evaluate the similarity between each entity pair between the two KGs. The features that have been used in entity embedding between two traditional KGs include in-KG context information (e.g. the semantics of OWL properties, co-occurrence of neighbors, compatible attribute values) and external information (e.g. external lexicons and Wikipedia links). For MMKGs, due to the introduction of multi-modal features, some entity-alignment

oriented MMKG embedding models are proposed [146], [147]. Feature vectors are encoded for different modalities respectively and then merged into one to represent the entity by the knowledge fusion techniques mentioned in the beginning of this subsection. One work [146] uses ranking loss as the loss function, while another [147] designs a loss function  $L = \alpha\|e - e_s\| + \beta\|e - e_n\| + \gamma\|e - e_i\|$  to enhance the complementarity of multiple modalities, where  $e_s, e_n, e_i$  is the embedding of three different modalities respectively  $e$  is the final embedding of the entity, and  $\alpha, \beta, \gamma$  is ratio hyper-parameters for each modality.

Another line of work [24] elaborates a Product of Experts (PoE) model to answer queries such as  $(h?, sameAs, t)$  or  $(h, sameAs, t?)$  where  $h$  and  $t$  are from different KGs. By incorporating [148] and extending it to visual features, the end-to-end learning framework has found to be superior to the concatenation and an ensemble type of approach as for entity alignment.

## 4.2 Out-of-MMKG Applications

The out-of-KG applications refer to the downstream applications that are not limited within the boundary of MMKGs, but could be assisted by them. In the following, we introduce several such applications as examples, including multi-modal named entity recognition and entity linking, visual question answering, image-text matching, multi-modal generation and multi-modal recommender system. Instead of providing a systematic reviews to all the solutions of these tasks, we mainly focus on introducing how MMKGs are utilized.

### 4.2.1 Multi-modal Entity Recognition and Linking

Named entity recognition (NER) with plain texts has been studied extensively. Ambiguity and diversity of entity mentions have always been the key challenges. Recent work considers to do NER to detect entities from texts attached with images, defined as multi-modal NER (MNER) [57], [117], where the images could provide necessary complementary information for entity recognition. MMKG plays an important role in MNER by providing vision features to describe different types of entities, such that the MNER models could better use the vision features of the images attached with the text for entity recognition. For instance, [149] proposes to use the background knowledge of images in MMKGs to help capture deep features of image to avoid error from shallow features.

Given a text with images attached, Multi-modal entity linking (MEL) uses both textual and visual information to map an ambiguous mention in the text to an entity in a knowledge base (KB) [150]. Although some early efforts does MEL based on a traditional KG as the KB, more and more recent work prefers to use MMKGs as the KB for linking. The knowledge with images in a MMKG are utilized by MEL in two ways: (1) Providing the target entities to which the entity mentions should be linked; (2) Learning distributed representations for each entity with multi-modal data, which are then used to measure the correlation between a mention and an entity. The usage of visual information with images would help to capture the relationship among mentions and entities [150], [151],

but the irrelevant part with images may also become noises and bring negative impact to the representation learning for both mentions and entities. To remove the side effect, a two-stage image and text correlation mechanism is proposed to filter out the irrelevant images based on the predefined threshold, and the multiple attention mechanisms is also utilized to capture the important information in the mention representation and entity representation by querying multi-hop entities around the mention’s candidate entities [118].

### 4.2.2 Visual Question Answering

Visual question answering (VQA) is a challenging task, which not only requires accurate semantic parsing to the questions, but also needs in-depth understanding to the correlations between different objects and scenes in the given image. In most of the recent VQA benchmark datasets such as GQA [119], OK-VQA [56] and KVQA [121], many questions require visual reasoning combined with external knowledge. The newly proposed VQA tasks bridge the discrepancy that humans can easily combine knowledge from various modalities to answer visual queries. For example, in the question “*Which American President is associated with the stuffed animal seen here?*”, if the stuffed animal in the image is detected as “*Teddy Bear*”, the answer inferred through KG will be “*Theodore Roosevelt*”, who is often referred as “*Teddy Roosevelt*”, and after whom Teddy Bear is named [56].

Extracting relationships between visual concepts and understanding semantic information in questions are two key issues for VQA. However, without incorporating more knowledge of various modalities, it is not scalable to reason just over image-question-answer triplets by semantic parsing and matching and can hardly generalize to more complicated cases [122]. MMKG helps to deal with the problems and enhances the interpretability to answers. First, MMKG provides knowledge about the named entities and their relations in the image which leads to deeper visual content understanding. Second, the structured symbolic knowledge in MMKG makes it a more explicit way to conduct the reasoning process and predict the final answers.

Some recent efforts [152] combines diverse KGs with different kinds of knowledge for VQA, including both unimodal KGs such as DBpedia [6], ConceptNet [2] and has-Part KB [153] for categorical and commonsense knowledge, and a MMKG Visual Genome [59] for visual data. Visual symbolic information in MMKG with graph-structured information conveying relations between visual concepts provides strong evidence to reason about the questions over graph network. Besides, the explicit semantic knowledge preserved in MMKG help refine the answers with more interpretability and generality [154]. The representations of different modalities preserved and unified in MMKG greatly benefit for relational reasoning across modalities.

### 4.2.3 Image-Text Matching

Image-text matching is a fundamental task in many text and image relevant applications such as image-text and text-image retrieval, which aims to output a semantic similarity score between the input image and text pair [114], [155], [156], [157], [158].

Image-text matching is usually achieved by mapping texts and images into a joint semantic space, and then learning a unified multi-modal representations for the similarity calculation. A general method is to exploit a multi-label detection module to extract semantic concepts and then fuse these concepts with global context of image [114], [159], [160]. However, it is difficult for pretrained detected-based models to find long-tail concepts, which constrains models to those detected concepts and leads to poor performance.

To overcome the bias in the training data for retrieval tasks, MMKG could be leveraged to expand more visual and semantic concepts leveraging the relations between multi-modal entities. Besides, MMKG can also help to construct scene graphs, which introduce informative correlation knowledge between visual concepts and further enhance image representations. For example, the concept pairs that frequently co-occurred in the multimodal triples of a MMKG, such as house-window and tree-leaf, can be extracted to enhance the representation of concepts in images, thus providing strong context signal for semantic understanding of images and leads to improved performance of image-text matching [16]. Besides, considering that one key step in image-text matching task is to align both local and global representations across different modalities, some efforts proposes to incorporate relations in MMKGs to represent both image and text with higher-level semantics [161]. Such graph-structured information better enhances the reasoning and inference capabilities of multi-modal data with more interpretability. MMKG also helps cross-modal alignment by learning a more unified multimodal representation.

#### 4.2.4 Multi-modal Generation Tasks

There are several studied multi-modal generation tasks such as image tagging, image captioning, visual storytelling etc. that could benefit from MMKGs.

**Image Tagging.** Traditional image tagging methods are easily limited to tag statistics bias as well as noisy and imprecise tags. By embedding the concept knowledge in MMKGs into images, the representations of images could be greatly improved [162], which as a result can enhance the performance of image tagging. Another effort [162] proposes to construct a MMKG named Visio-Textual Knowledge Base (VTKB), which includes textual and visual information as well as the relations between them. Based on this VTKB, a novel image tagging framework is proposed, incorporating visual information in VTKB to help disambiguate the concept and relate them better to images.

**Image Captioning.** The mainstream statistic-based image captioning models have two weaknesses: Firstly, they heavily rely on the performance of object detectors. The encoder-decoder framework with separate procedures of detection and captioning always leads to semantic inconsistency between the pre-defined objects/relations and target textual descriptions. Secondly, unseen objects always pose great challenges to them. The models trained on image-caption parallel corpora always fail to describe unseen objects and concepts.

Fortunately, MMKGs could help to alleviate the two obstacles of image captioning in the following ways: 1) Some efforts [163] proposes to leverage MMKG for relational

reasoning which results in more accurate and reasonable captions. More specifically, a semantic graph could be built for visual and knowledge vectors embedded from candidate image proposals, and the semantic graph could then be encoded for textual description generation. In this way, the semantic constraints summarized in MMKGs can be fully used, which may further endow the MMKGs ability and readily extended for more advanced reasoning. 2) The symbolic knowledge from MMKGs may enable the understanding of unseen objects [55]. Specifically, symbolic knowledge provides symbolic information about unseen objects and establishes semantic relation between seen objects and unseen objects in terms of symbolic knowledge. In the knowledge guided image-caption task containing novel objects, the key module is a multi-label image classifier for grounding depicted visual objects to knowledge base entities, unveiling a way to build a connection between real-world objects to their multi-modal information with the assistance of MMKGs [55]. By introducing external knowledge from MMKG-based multi-label classifier, image representations are expanded as well.

A more complex task named entity-aware image captioning, which asks for more informative descriptions of named entities given the background knowledge in the associated article. Though some studies extract and encode textual knowledge to construct a more fine-grained attention mechanism, they neglect the associations between named entities and visual cues in the image and therefore perform badly under some complicate scenarios. However, MMKGs can capture the fine-grained relationships between entities in the context and objects in the image, for generating captions with more accurate named entities and more relevant events [18]. More specifically, two different MMKGs are leveraged for various functions of different modules. First, in the cross-modal entity linking module, a complete MMKG is constructed by connecting a text subgraph and an image sub-graph extracted from the input article and image respectively, while in the meantime incorporating an external MMKG as an assistance. The well-established MMKG then together with the image and article greatly benefits the entity-aware caption generation procedure afterwards.

**Visual Storytelling.** Visual storytelling is a more challenging task than image captioning, which aims to tell the story according to a number of successive images. This task requires to discover the relations between the images and the objects associated with the images. Traditional visual storytelling approaches usually treat the task as a sequential image captioning problem and ignore the relation between image, which may produce monotonous stories. Besides, these approaches are limited to the vocabulary and knowledge in a single training dataset. To tackle these problems, some recent efforts [164] resorts to a MMKG for help within a distill-enrich-generate three stage framework. After first extracting a set of words from each image, all terms from two consecutive images are paired to query Visual Genome for object relations and OpenIE for term relations, for all possible tuples and generate additional story sentences. The most reasonable term sets are then selected and fed to next story generation step. In this way of mimicking how humans generate a story, the usage of relations in knowledge

graphs provide a strong logical inference between images, making the generated story more fluent.

#### 4.2.5 Multi-modal Recommender System

Recommender system aims to recommend items that users might like/buy through the analysis of historical data. Various factors needs to be balanced during the process, such as accuracy, novelty, dispersity and stability [165]. Where there are multi-modal data such as image and text in a recommending scenario, we say it is a multi-modal recommender system, where the information of different modals should be leveraged jointly.

Recent years have proved that MMKGs could greatly enhance multi-modal recommender system [166]. Some approaches obtain the representations with rich semantics for items by leveraging external MMKGs. Incorporating information of MMKG across different modalities can help solve the cold-start problem long existing in Collaborative Filtering (CF) based recommending strategies [167]. Some other approaches find other ways to utilize MMKG for more personalized and explainable recommendations. For instance, [168] fully exploits the graph structure of MMKG and designs a novel approach hierarchy attention-path over MMKGs for the reasoning over items with information across different modalities. Rich path semantics could be learned through entities and images within the path in MMKG, thus producing an interpretable and explicit recommendation with higher knowledge level. Differently, some recent efforts [131], [168] novelly proposes to construct a personalized MMKG from the images and texts of items in various ways, and then the entity relation reasoning between items can be better modeled by taking the relations in MMKG into account.

## 5 OPEN PROBLEMS

This section discusses some open problems on MMKG construction and application leaving for future research.

### 5.1 Complex Symbolic Knowledge Grounding

Besides the grounding of entities, concepts and relations, some downstream applications also require the grounding of complex symbolic knowledge which consists of multiple relational facts that have close semantic relations with each other. These multiple relational facts may be a path or a subgraph in the KG. For example, for a subgraph in KG containing Trump's wife, daughter, grandson etc., a proper grounding image might be a Trump's family photo. This motivates *multiple relational grounding*, which aims to find images to express the knowledge contained in a path or a subgraph in KG.

Multiple relational grounding is challenging since it involves the grounding of more than one relations and these multiple groundings are usually interleaved with each other in a complicated way. We have to find the images that fully embody the composite semantic relations. The composite semantic in many cases is only implicit expressed and might change over time.

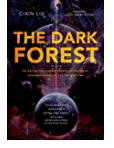
entity	Pluvianus aegyptius	The Wandering Earth	arrogance
image			

TABLE 6: Examples of quality problems in MMKG.

### 5.2 Quality Control

In general, we rely on data-driven approaches to build a large scale MMKG. A MMKG that automatically harvested from big data inevitably suffer from quality issues, that is, the MMKG might contain errors, missing facts or outdated facts. For example, in search behavior data based multi-modal knowledge acquisition, it is easy to associate a wrong image for a long-tail entity because such entity might have no image on Web, thus any clicked image are misleading to a mistakenly grounding.

Besides the common quality problems in accuracy, completeness, consistency and freshness, which are discussed and studied extensively in traditional KG, MMKG has some special quality issues that concerns about the images. Firstly, the image of some entity might be easily mixed with another when the two entities are closely-related with each other. See the first example in Table 6, Pluvianus aegyptius is a kind of bird that has a symbiosis with crocodiles, so we may always get a picture of both crocodile and the bird when searching for it. Secondly, the images of more famous entity may easily appear in the entity grounding results of its closely-related entities. See the second example in Table 6, The Wandering Earth is wrote by the famous Chinese science fiction writer Liu Cixin. While searching for this book, we can always get a picture of another more famous book named The dark forest of him. Thirdly, there are some abstract concepts whose visual features are not clear enough. For example, visual features of the abstract noun arrogance are unfixed, so we can always get some pictures that are completely irrelevant. To tackle the above problems, more visual analysis and background information might be needed to guide click-through rates and text information to avoid this misdirection.

### 5.3 Efficiency

Efficiency is always an non-negligible issue when building a large scale KG. The efficiency problem of constructing a MMKG is more striking, since the extra complexity to process multimedia data needs to be considered. For example, it takes NEIL [19] around 350K CPU hours to collect 400K visual instances for 2273 objects, while in a typical KG we need to ground billions of instances. The scalability of the existing solutions in building MMKGs will be greatly challenged. If the grounding objective is video data, the scalability issue might be even amplified.

Besides the construction of MMKG, the online application of MMKG also needs to carefully address the efficiency issue since the MMKG needs to serve applications in real time. The efficiency of the solution is crucial for the building an online MMKG-based applications.

## 6 CONCLUSION

We are the first to thoroughly survey the existing work on MMKGs constructed by texts and images. We systematically review the existing work on MMKG construction and application. We compare mainstream MMKGs in terms of what they contain and how they construct. We analyse the strengthens and weakness of different solutions in MMKG construction and applications. We not only point out some potential opportunities with the existing tasks in both MMKG construction and application, but also list some promising future directions with the construction and application of MMKGs.

## REFERENCES

- [1] C. Matuszek, M. Witbrock, J. Cabral, and J. DeOliveira, "An introduction to the syntax and content of cyc," *UMBC Computer Science and Electrical Engineering Department Collection*, 2006.
- [2] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.
- [3] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [4] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network," in *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 2010, pp. 216–225.
- [5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.
- [6] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.
- [7] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 697–706.
- [8] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.
- [9] B. Xu, Y. Xu, J. Liang, C. Xie, B. Liang, W. Cui, and Y. Xiao, "Cn-dbpedia: A never-ending chinese knowledge extraction system," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2017, pp. 428–438.
- [10] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probbase: A probabilistic taxonomy for text understanding," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, 2012, pp. 481–492.
- [11] M. Wick and B. Vatant, "The geonames geographical database," *Available from World Wide Web: http://geonames.org*, 2012.
- [12] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.
- [13] ——, "Symbol grounding problem," *Encyclopedia of cognitive science*, 2003.
- [14] L. Steels, "The symbol grounding problem has been solved. so what's next," *Symbols and embodiment: Debates on meaning and cognition*, pp. 223–244, 2008.
- [15] E. M. Bender and A. Koller, "Climbing towards nlu: On meaning, form, and understanding in the age of data," in *Proc. of ACL*, 2020.
- [16] B. Shi, L. Ji, P. Lu, Z. Niu, and N. Duan, "Knowledge aware semantic concept expansion for image-text matching," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019, pp. 5182–5189.
- [17] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12700–12710.
- [18] W. Zhao, Y. Hu, H. Wang, X. Wu, and J. Luo, "Boosting entity-aware image captioning with multi-modal knowledge graph," *arXiv preprint arXiv:2107.11970*, 2021.
- [19] X. Chen, A. Shrivastava, and A. Gupta, "Neil: Extracting visual knowledge from web data," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1409–1416.
- [20] M. Li, A. Zareian, Y. Lin, X. Pan, S. Whitehead, B. Chen, B. Wu, H. Ji, S.-F. Chang, C. Voss et al., "Gaia: A fine-grained multimedia knowledge extraction system," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020, pp. 77–86.
- [21] H. Wen, Y. Lin, T. Lai, X. Pan, S. Li, X. Lin, B. Zhou, M. Li, H. Wang, H. Zhang et al., "Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, 2021, pp. 133–143.
- [22] S. Ferrada, B. Bustos, and A. Hogan, "Imgpedia: a linked dataset with content-based analysis of wikimedia images," in *International Semantic Web Conference*. Springer, 2017, pp. 84–93.
- [23] D. Oñoro-Rubio, M. Niepert, A. García-Durán, R. González, and R. J. López-Sastre, "Answering visual-relational queries in web-extracted knowledge graphs," *arXiv preprint arXiv:1709.02314*, 2017.
- [24] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio, and D. S. Rosenblum, "Mmk: Multi-modal knowledge graphs," in *European Semantic Web Conference*. Springer, 2019, pp. 459–474.
- [25] M. Wang, H. Wang, G. Qi, and Q. Zheng, "Richpedia: a large-scale, comprehensive multi-modal knowledge graph," *Big Data Research*, vol. 22, p. 100159, 2020.
- [26] H. Alberts, T. Huang, Y. Deshpande, Y. Liu, K. Cho, C. Vania, and I. Calixto, "Visualsem: a high-quality knowledge graph for vision and language," *arXiv preprint arXiv:2008.09150*, 2020.
- [27] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.
- [28] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8307–8316.
- [29] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.
- [30] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.
- [31] Z. Yang, B. Gong, L. Wang, W. Huang, D. Yu, and J. Luo, "A fast and accurate one-stage approach to visual grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4683–4693.
- [32] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.
- [33] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "Fvqa: Fact-based visual question answering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2413–2427, 2018.
- [34] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 35–44.
- [35] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5764–5773.
- [36] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 394–10 403.
- [37] A. Karpathy, A. Joulin, and L. Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *arXiv preprint arXiv:1406.5679*, 2014.
- [38] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.
- [39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [41] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.
- [42] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep visual-semantic hashing for cross-modal retrieval," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1445–1454.
- [43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
- [44] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
- [45] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, vol. 34, no. 6, pp. 96–108, 2017.
- [46] T. Zhu, Y. Wang, H. Li, Y. Wu, X. He, and B. Zhou, "Multimodal joint attribute prediction and value extraction for e-commerce product," *arXiv preprint arXiv:2009.07162*, 2020.
- [47] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.
- [48] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Learning universal image-text representations," *arXiv preprint arXiv:1909.11740*, 2019.
- [49] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11336–11344.
- [50] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," *arXiv preprint arXiv:2102.03334*, 2021.
- [51] W. Li, C. Gao, G. Niu, X. Xiao, H. Liu, J. Liu, H. Wu, and H. Wang, "Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning," *arXiv preprint arXiv:2012.15409*, 2020.
- [52] 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," *arXiv preprint arXiv:2006.16934*, vol. 1, p. 12, 2020.
- [53] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "ViLbert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.
- [54] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," *arXiv preprint arXiv:1908.07490*, 2019.
- [55] A. Mogadala, U. Bista, L. Xie, and A. Rettinger, "Describing natural images containing novel objects with knowledge guided assistance," *arXiv preprint arXiv:1710.06303*, 2017.
- [56] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3195–3204.
- [57] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive co-attention network for named entity recognition in tweets," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [58] P. Perona, "Vision of a visipedia," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1526–1534, 2010.
- [59] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.
- [60] K. Chen, J. Gao, and R. Nevatia, "Knowledge aided consistency for weakly supervised phrase grounding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4042–4050.
- [61] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7206–7215.
- [62] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [63] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
- [64] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.
- [65] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J.Pont-Tuset, S. Kamali, S. Popov, M. Mallochi, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *IJCV*, 2020.
- [66] M. Wang, M. Azab, N. Kojima, R. Mihalcea, and J. Deng, "Structured matching for phrase localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 696–711.
- [67] M. Li, A. Zareian, Q. Zeng, S. Whitehead, D. Lu, H. Ji, and S.-F. Chang, "Cross-media structured common space for multimedia event extraction," *arXiv preprint arXiv:2005.02472*, 2020.
- [68] 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," *arXiv preprint arXiv:2103.00020*, 2021.
- [69] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clip-score: A reference-free evaluation metric for image captioning," *arXiv preprint arXiv:2104.08718*, 2021.
- [70] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021.
- [71] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.
- [72] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1292–1302.
- [73] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 9–16.
- [74] ———, "Modeling mutual context of object and human pose in human-object interaction activities," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 17–24.
- [75] S. Maji, L. Bourdev, and J. Malik, "Action recognition from a distributed representation of pose and appearance," in *CVPR 2011*. IEEE, 2011, pp. 3177–3184.
- [76] S. Antol, C. L. Zitnick, and D. Parikh, "Zero-shot learning via visual abstraction," in *European conference on computer vision*. Springer, 2014, pp. 401–416.
- [77] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with \*cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1080–1088.
- [78] M. Elhoseiny, S. Cohen, W. Chang, B. Price, and A. Elgammal, "Sherlock: Scalable fact learning in images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
- [79] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5831–5840.
- [80] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *European conference on computer vision*. Springer, 2016, pp. 852–869.
- [81] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE conference on computer vision and Pattern recognition*, 2017, pp. 3076–3086.
- [82] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proceed-*

- ings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5532–5540.
- [83] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, “Scene graph generation by iterative message passing,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5410–5419.
- [84] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, “Graph r-cnn for scene graph generation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–685.
- [85] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, “Learning to compose dynamic tree structures for visual contexts,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6619–6628.
- [86] W. Wang, R. Liu, M. Wang, S. Wang, X. Chang, and Y. Chen, “Memory-based network for scene graph with unbalanced relations,” in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2400–2408.
- [87] W. Wang, M. Wang, S. Wang, G. Long, L. Yao, G. Qi, and Y. Chen, “One-shot learning for long-tail visual relation detection,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12225–12232.
- [88] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, “Large-scale visual relationship understanding,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9185–9194.
- [89] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, “Unbiased scene graph generation from biased training,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3716–3725.
- [90] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, “Graphical contrastive losses for scene graph parsing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11535–11543.
- [91] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, M. Chen, Z. Ma, S. Wang, H.-S. Fang, and C. Lu, “Hake: Human activity knowledge engine,” *arXiv preprint arXiv:1904.06539*, 2019.
- [92] T. Zhang, S. Whitehead, H. Zhang, H. Li, J. Ellis, L. Huang, W. Liu, H. Ji, and S.-F. Chang, “Improving event extraction via multimodal integration,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 270–278.
- [93] B. Chen, X. Lin, C. Thomas, M. Li, S. Yoshida, L. Chum, H. Ji, and S.-F. Chang, “Joint multimedia event extraction from video and article,” *arXiv preprint arXiv:2109.12776*, 2021.
- [94] M. Yatskar, L. Zettlemoyer, and A. Farhadi, “Situation recognition: Visual semantic role labeling for image understanding,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5534–5542.
- [95] S. Pratt, M. Yatskar, L. Weih, A. Farhadi, and A. Kembhavi, “Grounded situation recognition,” in *European Conference on Computer Vision*. Springer, 2020, pp. 314–332.
- [96] Y. Xiong, K. Zhu, D. Lin, and X. Tang, “Recognize complex events from static images by fusing deep channels,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1600–1609.
- [97] H. Li, J. G. Ellis, H. Ji, and S.-F. Chang, “Event specific multimodal pattern mining for knowledge base construction,” in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 821–830.
- [98] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, “Abstract meaning representation for sembanking,” in *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, 2013, pp. 178–186.
- [99] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [100] X. Zhang, X. Sun, C. Xie, and B. Lun, “From vision to content: Construction of domain-specific multi-modal knowledge graph,” *IEEE Access*, vol. 7, pp. 108278–108294, 2019.
- [101] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, “Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 547–558.
- [102] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.
- [103] Y. Yan, G. Liu, S. Wang, J. Zhang, and K. Zheng, “Graph-based clustering and ranking for diversified image search,” *Multimedia Systems*, vol. 23, no. 1, pp. 41–52, 2017.
- [104] S. K. Divvala, A. Farhadi, and C. Guestrin, “Learning everything about anything: Webly-supervised visual concept learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3270–3277.
- [105] Q. Mei-Bin, Z. Jun-Jun, J. Ping, and J. Jian-Guo, “Representative image selection from image dataset,” *ACTA AUTOMATICA SINICA*, vol. 40, no. 4, pp. 706–712, 2014.
- [106] H. Yu, Z.-H. Deng, Y. Yang, and T. Xiong, “A joint optimization model for image summarization based on image content and tags,” in *Twenty-eighth AAAI conference on artificial intelligence*, 2014.
- [107] Y. Wang, L. Zhu, and X. Qian, “Social image retrieval based on topic diversity,” *Multimedia Tools and Applications*, vol. 80, no. 8, pp. 12367–12387, 2021.
- [108] R. H. Van Leuken, L. Garcia, X. Olivares, and R. van Zwol, “Visual diversification of image search results,” in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 341–350.
- [109] M. Wang, K. Yang, X.-S. Hua, and H.-J. Zhang, “Towards a relevant and diverse search of social images,” *IEEE Transactions on Multimedia*, vol. 12, no. 8, pp. 829–842, 2010.
- [110] T. Deselaers, T. Gass, P. Drewes, and H. Ney, “Jointly optimising relevance and diversity in image retrieval,” in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–8.
- [111] Z. Ji, Y. Su, Y. Pang, and X. Qu, “Diversifying the image relevance reranking with absorbing random walks,” in *2011 Sixth International Conference on Image and Graphics*. IEEE, 2011, pp. 981–986.
- [112] R. Raguram and S. Lazebnik, “Computing iconic summaries of general visual concepts,” in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2008, pp. 1–8.
- [113] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, “Image retrieval using scene graphs,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.
- [114] Y. Huang, Q. Wu, C. Song, and L. Wang, “Learning semantic concepts and order for image and sentence matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6163–6171.
- [115] Y. Guo, J. Chen, H. Zhang, and Y.-G. Jiang, “Visual relations augmented cross-modal retrieval,” in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 9–15.
- [116] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, “Cross-modal scene graph matching for relationship-aware image-text retrieval,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1508–1517.
- [117] S. Moon, L. Neves, and V. Carvalho, “Zeroshot multimodal named entity disambiguation for noisy social media posts,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018.
- [118] L. Zhang, Z. Li, and Q. Yang, “Attention-based multimodal entity linking with high-quality images,” in *International Conference on Database Systems for Advanced Applications*. Springer, 2021, pp. 533–548.
- [119] D. A. Hudson and C. D. Manning, “Gqa: A new dataset for real-world visual reasoning and compositional question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6700–6709.
- [120] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, “Visual7w: Grounded question answering in images,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.
- [121] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, “Kvqa: Knowledge-aware visual question answering,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8876–8884.
- [122] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, “Explicit knowledge-based reasoning for visual question answering,” *arXiv preprint arXiv:1511.02570*, 2015.
- [123] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “Nuswide: a real-world web image database from national university

- of singapore," in *Proceedings of the ACM international conference on image and video retrieval*, 2009, pp. 1–9.
- [124] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 39–43.
- [125] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko, "Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2712–2719.
- [126] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12466–12475.
- [127] A. Tran, A. Mathews, and L. Xie, "Transform and tell: Entity-aware news image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13035–13045.
- [128] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra *et al.*, "Visual storytelling," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1233–1239.
- [129] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *AcM transactions on interactive intelligent systems (tiis)*, vol. 5, no. 4, pp. 1–19, 2015.
- [130] A. Uyar and F. M. Aliyu, "Evaluating search features of google knowledge graph and bing satori: entity types, list searches and query interfaces," *Online Information Review*, 2015.
- [131] R. Sun, X. Cao, Y. Zhao, J. Wan, K. Zhou, F. Zhang, Z. Wang, and K. Zheng, "Multi-modal knowledge graphs for recommender systems," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 1405–1414.
- [132] Y. Huang, M. Li, and Y. Wu, "Kkbox's music recommendation."
- [133] A. Bordes, J. Weston, R. Collobert, and Y. Bengio, "Learning structured embeddings of knowledge bases," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.
- [134] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.
- [135] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.
- [136] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.
- [137] G. Ji, S. He, L. Xu, K. Liu, and J. Zhao, "Knowledge graph embedding via dynamic mapping matrix," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 687–696.
- [138] H. Mousselly-Sergieh, T. Botschen, I. Gurevych, and S. Roth, "A multimodal translation-based approach for knowledge graph representation learning," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 2018, pp. 225–234.
- [139] A. Rettinger, "Towards holistic concept representations: Embedding relational knowledge, visual attributes, and distributional word semantics," 2017.
- [140] P. Pezeshkpour, L. Chen, and S. Singh, "Embedding multimodal relational data for knowledge base completion."
- [141] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun, "Representation learning of knowledge graphs with entity descriptions," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [142] Y. Lin, Z. Liu, H. Luan, M. Sun, S. Rao, and S. Liu, "Modeling relation paths for representation learning of knowledge bases," *arXiv preprint arXiv:1506.00379*, 2015.
- [143] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in neural information processing systems*, 2013, pp. 926–934.
- [144] R. Xie, Z. Liu, H. Luan, and M. Sun, "Image-embodied knowledge representation learning," *arXiv preprint arXiv:1609.07028*, 2016.
- [145] W. Wilcke, P. Bloem, V. de Boer, R. van t Veer, and F. van Harmelen, "End-to-end entity classification on multimodal knowledge graphs," *arXiv preprint arXiv:2003.12383*, 2020.
- [146] H. Guo, J. Tang, W. Zeng, X. Zhao, and L. Liu, "Multimodal entity alignment in hyperbolic space," *arXiv preprint arXiv:2106.03619*, 2021.
- [147] L. Chen, Z. Li, Y. Wang, T. Xu, Z. Wang, and E. Chen, "Mmea: Entity alignment for multi-modal knowledge graph," in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2020, pp. 134–147.
- [148] A. Garcia-Duran and M. Niepert, "Kblrn: End-to-end learning of knowledge base representations with latent, relational, and numerical features," *arXiv preprint arXiv:1709.04676*, 2017.
- [149] D. Chen, Z. Li, B. Gu, and Z. Chen, "Multimodal named entity recognition with image attributes and image knowledge," in *International Conference on Database Systems for Advanced Applications*. Springer, 2021, pp. 186–201.
- [150] O. Adjali, R. Besançon, O. Ferret, H. Le Borgne, and B. Grau, "Multimodal entity linking for tweets," *Advances in Information Retrieval*, vol. 12035, p. 463, 2020.
- [151] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity disambiguation for noisy social media posts," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2000–2008.
- [152] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, "Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14111–14121.
- [153] S. Bhakthavatsalam, K. Richardson, N. Tandon, and P. Clark, "Do dogs have whiskers? a new knowledge base of haspart relations," *arXiv preprint arXiv:2006.07510*, 2020.
- [154] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu, and J. Tan, "Cross-modal knowledge reasoning for knowledge-based visual question answering," *Pattern Recognition*, vol. 108, p. 107563, 2020.
- [155] L. Ma, Z. Lu, L. Shang, and H. Li, "Multimodal convolutional neural networks for matching image and sentence," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2623–2631.
- [156] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.
- [157] L. Ma, W. Jiang, Z. Jie, Y.-G. Jiang, and W. Liu, "Matching image and sentence with multi-faceted representations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2250–2261, 2019.
- [158] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020.
- [159] Y. Huang, Q. Wu, W. Wang, and L. Wang, "Image and sentence matching via semantic concepts and order learning," *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [160] W. Wang, Y. Huang, and L. Wang, "Language-driven temporal activity localization: A semantic matching reinforcement learning model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 334–343.
- [161] H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *European Conference on Computer Vision*. Springer, 2020, pp. 18–34.
- [162] C. Chaudhary, P. Goyal, D. N. Prasad, and Y.-P. P. Chen, "Enhancing the quality of image tagging using a visio-textual knowledge base," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 897–911, 2019.
- [163] J. Hou, X. Wu, Y. Qi, W. Zhao, J. Luo, and Y. Jia, "Relational reasoning using prior knowledge for visual captioning," *arXiv preprint arXiv:1906.01290*, 2019.
- [164] C.-C. Hsu, Z.-Y. Chen, C.-Y. Hsu, C.-C. Li, T.-Y. Lin, T.-H. Huang, and L.-W. Ku, "Knowledge-enriched visual storytelling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 7952–7960.
- [165] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-based systems*, vol. 46, pp. 109–132, 2013.
- [166] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, 2009.
- [167] F. Zhang, N. J. Yuan, D. Lian, X. Xie, and W.-Y. Ma, "Collaborative knowledge base embedding for recommender systems," in

- Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 353–362.
- [168] S. Tao, R. Qiu, Y. Ping, and H. Ma, “Multi-modal knowledge-aware reinforcement learning network for explainable recommendation,” *Knowledge-Based Systems*, p. 107217, 2021.