Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Prospect:

Visual knowledge in the big model era: retrospect and prospect*

Wenguan WANG, Yi YANG‡, Yunhe PAN

College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China

E-mail: wenguanwang@zju.edu.cn; yiyang@zju.edu.cn; panyh@zju.edu.cn

Received Apr. 2, 2024; Revision accepted May 26, 2024; Crosschecked

Abstract: Visual knowledge is a new form of knowledge representation that can encapsulate visual concepts and their relations in a succinct, comprehensive, and interpretable manner, with a deep root in cognitive psychology. As the knowledge of the visual world has been identified as an indispensable component of human cognition and intelligence, visual knowledge is poised to have a pivotal role in establishing machine intelligence. With the recent advance of Artificial Intelligence (AI) techniques, large AI models (or foundation models) have emerged as a potent tool capable of extracting versatile patterns from broad data as implicit knowledge, and abstracting them into an outrageous amount of numeric parameters. To pave the way for creating visual knowledge empowered AI machines in this coming wave, we present a timely review that investigates the origins and development of visual knowledge in the pre-big model era, and accentuates the opportunities and unique role of visual knowledge in the big model era.

Key words: Visual knowledge; Artificial intelligence; Foundation model; Deep learning
https://doi.org/10.1631/FITEE.2400250 **CLC number:**

unedited

1 Introduction

The concept of visual knowledge (Pan, 2019) was recently proposed as a new form of knowledge representation that differs from the traditional ones used or learned by symbolic and sub-symbolic AI approaches (*e.g.*, knowledge graph, handcrafted image descriptors, and distributed visual representations). Drawing on cognitive studies (Anderson, 2005) of human mental imagery which enables us to manipulate visual entities in our mind, the visual knowledge theory posits that next-generation AI needs to fully express visual concepts and their attributes (*e.g.*, shape, structure, motion, and affordance), as well as reason about their transformations, compositions, comparisons, predictions, and narrations, through a unified, abstract, and interpretable form of representation.

After the emergence of large language models like GPT-3 (Brown et al., 2020), the field of natural language processing has experienced remarkable advancements: traditional "narrow" language models that are trained to perform specific tasks in a single domain are giving way to highly sophisticated and versatile language models that are trained on a vast corpus of unlabeled textual data that can be used for different language tasks across domains. Like GPT for natural language processing, the recent work known as Segment Anything Model (SAM) (Kirillov et al., 2023) ushered the field of computer vision into the era of visual foundation models — by training on >1B segmentation masks in >11M natural images, SAM shows the promise of a broad applicability to various image segmentation tasks, without re-training or fine-tuning as previously needed. With incredible

‡ Corresponding author
* This project supported by the National Science and Technology Major Project (No. 2023ZD0121300), the National Natural Science Foundation of China (No. 62372405), and the Fundamental Research Funds for the Central Universities
ORCID: Wenguan WANG, https://orcid.org/0000-0002-0802-9567
© Zhejiang University Press 2024

speed, large models are revolutionizing AI field and transforming the landscape of scientific research.

Albeit the unprecedented progress, it is becoming increasingly evident that large AI models still suffer several deficiencies that compromise their reliability and efficacy. Chief among these is their pronounced opacity, which poses great challenges for trust, accountability, and effective debugging, as well as their insatiable demand for data and computational resources, leading to both ethical and environmental concerns. These limitations are inherited from their rudimentary predecessors but exacerbated by their heightened sophistication and scale. Compounding these concerns, large AI models are susceptible to generating nonsensical or unfaithful content, known as "hallucination", exposing their inherent biases, lack of real-world understanding, and weakness in generalizing or reasoning beyond their scope.

Considering the appealing advantages of visual knowledge in terms of expressive and interpretable representation, manipulation, and reasoning of visual concepts, it is probably fair to assume that a deeper understanding and development of visual knowledge can (at least in part) alleviate the weaknesses of big AI models. On the other hand, given the tremendous success of big AI models and the significant challenges of visual knowledge acquisition, it appears to be apparent that future endeavors should be made to develop specific techniques that seek to build visual knowledge with the aid of large-scale statistical learning. These are the considerations that give rise to the work presented here. In this work, we delve into early theoretical and methodological investigations of visual knowledge, and demonstrate that the key insights provided by visual knowledge studies shed new light onto the increasingly prominent role of trust, interpretability, and accountability in the ongoing AI revolution sparked by big models. Moreover, we identify promising directions for the creation of more powerful AI systems that harness the synergies between visual knowledge and big models to overcome the weaknesses of each other. This also turns out that the development of new, deeper forms of visual knowledge in large-scale charts the course for next-generation AI (Pan, 2020; Yang et al., 2021).

The following sections are organized as follows (Fig. 1): we first briefly introduce the origin and theory of visual knowledge (Sec. 2). Subsequently, we systematically categorize the latest advances in visual knowledge based on a taxonomy from four different perspectives, including visual concept, visual relation, visual operation, and visual reasoning, which are key elements and characteristics of visual knowledge (Sec. 3). Based on the analysis of the research situation of visual knowledge, we outline a series of promising directions that can act as a compass for future explorations of visual knowledge in the imminent age of big models (Sec. 4). Sec. 5 concludes this work. We hope our efforts in this paper can bring together computer vision, graphics, and machine learning communities as well as the industry to advance a further development of visual knowledge, and inch the level of general intelligence in machines closer to that of humans.

2 Visual knowledge: origins and definitions

To facilitate a comprehensive understanding of visual knowledge theory, in this section, we discuss its cognitive foundations (Sec. 2.1) and core definitions (Sec. 2.2).

2.1 Origins

The theory of visual knowledge (Pan, 2019) did not appear "out of the blue." Rather, its roots extend deeply into the realm of cognitive psychology.

The significant role of visual signals in information processing of human brain

Our knowledge of the world is not solely derived from textual and verbal material, but also from the visual perception of the real environment. Pioneering research in cognitive and biological psychology (Milner and Goodale, 2006) has revealed that nearly half of our cerebral cortex is dedicated to processing visual stimuli. In addition, the human brain processes images 60,000 times faster than text, and 90 percent of information transmitted to the brain is visual. These statistics suggest that the human brain places a high value on visuals over any other type of information.

Visual memory: capacity, function, and representation of the stored content

Recent cognitive research also provided strong evidence that the human memory for pictures is much better than that for sounds (Bigelow

unedited

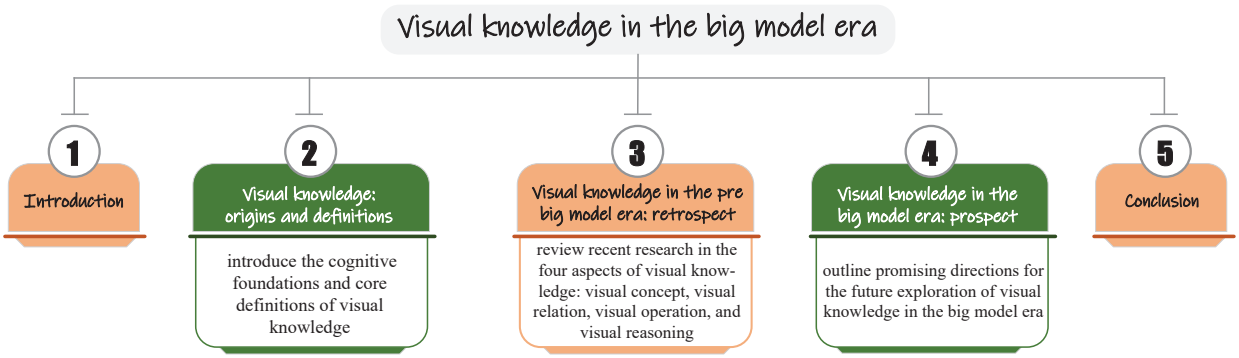

Fig. 1 Overall structure of this article.

and Poremba, 2014). Human *visual memory* is ubiquitous in daily life and closely links to many high-level cognitive functions, such as *mental imagery* (the ability to create an image in mind in the absence of sensory input). Visual memory, actively (*i.e.*, visual working memory) or passively (*i.e.*, visual long-term memory), holds and recalls visual information in mind, making it accessible and manipulable in support of ongoing cognitive tasks. Rather than solely concentrating on the capacity and cognitive function of visual memory, cognitive psychologists endeavored to explore the representation of the content stored in visual memory. Shepard and Metzler (1988); Shepard and Feng (1972); Kosslyn et al. (1978); Moyer (1973) conducted a series of experiments showing that visual memory representations, in contrast to verbal memory representations, support a variety of mental manipulations, including rotation, folding, scanning, and analogizing. Cognitive psychologists also found evidence suggesting that the structure of visual memory representations can be thought of as hierarchically organized (Brady et al., 2011).

Interactions among perception, visual memory, and human knowledge

Compared with verbal memory, which is primarily processed in the left hemisphere, visual memory tends to be more bilateral. Visual memory can be *episodic* (*i.e.*, memory of visual events or experiences that have a specific time and place), but it can also be *semantic* (*i.e.*, memory of general facts or visual concepts that are not tied to a specific context). Human store knowledge about most items in the real world and there is clear evidence that the representation of content in visual memory is not just a straightforward recording of sensory input but depends upon our past experience and our stored knowledge (Brady et al., 2011). Here it is necessary to make the distinction between visual memory and human stored knowledge. Visual memory refers to the ability to remember visual information that was seen previously. Thus, visual memory is the storage and subsequent retrieval of perceived visual information. Stored knowledge refers to the preexisting representations that underlie our ability to recognize and understand visual input. For example, when we first view an image, say of an orange, stored knowledge about the visual form and features of oranges in general enables us to recognize the object as such. Later, if we encounter another picture of an orange, visual memory enables us to decide whether it is the exact same orange we saw previously. Thus, specific items for which we have expertise, like faces, are represented with more fidelity (Scolari et al., 2008), while general concepts are represented after statistical regularities (Brady et al., 2008). Cognitive research has also shown that our stored knowledge can modulate how we form and utilize mental images and visual memory, and our knowledge and visual memory can affect how we perceive and attend to visual stimuli.

Proposal of the visual knowledge theory

The aforementioned cognitive studies evidenced the close and complex relations between visual information processing, visual memory, and human knowledge: the information gleaned from visual experiences supports many cognitive functions (such as visual memory and mental imagery) as well as the construction of our knowledge; and such knowledge, in turn, shapes visual memory, influences percep-

unedited

tion, and greatly facilitates our understanding of the world around us. Taking all these into consideration, a plausible argument can be made that one of the shortcomings of existing AI research is the scarcity of studies concerning human mental representations of visual items. The theory of visual knowledge (Pan, 2019) is thus arising to fill such gap.

2.2 Definitions

Basically, our visual knowledge is stable mental representations of visual objects and the commonalities in the inherent rules among various tasks. They are abstracted from our visual experiences and memory, and retained in our mind. They enable us to remember, imagine, and reason about the world and accomplish targeted tasks. Neuropsychological investigations have also revealed some characteristics of our mental representations of visual objects:

- The ability to capture the typical attributes of visual objects, such as their shapes, sizes, colors, and textures.
- The ability to describe the static and dynamic relationships between visual objects, such as relative positions, actions, velocities, and temporal sequences.
- The ability to perform spatio-temporal operations over visual objects, such as transforming shapes, actions, and scenes, making analogies and associations, and predicting future outcomes.
- The ability to engage in reasoning, such as analogizing, inducting, and deducing new tasks, combining existing concepts to form new concepts, and generalizing from anomalous samples.

Visual knowledge is not just an abstract representation of visual objects, but involves an active and generative process that supports various cognitive skills. Hence one of the core insights of the visual knowledge theory is that AI systems should also develop and use visual knowledge in a similar way.

More specifically, visual knowledge (Pan, 2019), as a new form of knowledge representation, is constructed as a combination of four essential components, namely visual concept (Sec. 2.2.1), visual relation (Sec. 2.2.2), visual operation (Sec. 2.2.3), and visual reasoning (Sec. 2.2.4). With these key components, visual knowledge can enable AI systems to comprehensively describe, robustly recognize, and reason about visual items and solve tasks.

2.2.1 Visual concept

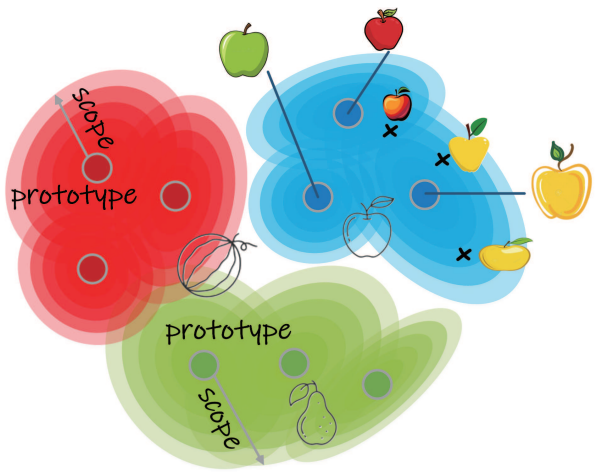

Fig. 2 Illustration of prototype-and scope-based visual concept representation. Here we show three visual concepts, namely *pear*, *apple*, and *watermelon*.

A visual concept is a category of visual objects that share some common features. The visual knowledge theory holds that a visual concept is defined by *prototype* and *scope*. When thinking of a visual concept, such as apple, we form a mental set of images that represent the most common or typical features/attributes of that concept. These images are called prototypes, serving as the basis for generating or recognizing any variation of that concept. For instance, we might have prototypes of apples that are red, green, or yellow, and those are round, oval, or heart-shaped. Based on these prototypes, we can imagine or identify any apple that has similar features, even if it is not exactly the same as any of the prototypes. But apples do not look exactly the same; some might be lighter or darker, bigger or smaller, smoother or rougher than others. However, there exists a limit or boundary to how much an apple can deviate from the prototypes and still be considered an apple. If the shape or color is too different, it might belong to another visual concept, such as *pear* or *watermelon*. The range of variation that is acceptable for a category is called scope. Shapes and colors within the scope are considered part of the category of apples, but shapes and colors outside it are not (see Fig. 2).

The idea of using prototypes to represent visual concepts is in line with the classic *prototype theory*, which in large part owes its beginnings to Rosch and

unedited

Mervis (1975), and has gained widespread recognition in cognitive science and other fields. The prototype theory provides an important theoretical account of cognitive categorization. It posits that a category of things in the world (objects, animals, shapes, *etc.*) can be represented in the mind by a prototype. A prototype is a cognitive representation that captures the regularities and commonalities among category members. According to the prototype theory, objects are classified by first comparing them to the prototypes stored in the memory, evaluating the similarity evidence from those comparisons, and then classifying the item in accord with the most similar prototype. Formally, let $\mathcal{X}$ be the data space and $\mathcal{Y}=\{y_1,\cdots,y_C\}$ a set of C categories. Given an data instance $\boldsymbol{x}\in\mathcal{X}$, a prototype classification model assigns it to the class $y\in\mathcal{Y}$ with the closest prototype:

$$y=\arg\min_{y_c\in\mathcal{Y}}\langle\boldsymbol{x},\boldsymbol{p}_c\rangle, \tag{1}$$

where $\langle\cdot,\cdot\rangle$ is a distance measure, $\boldsymbol{p}_c$ refers to the prototype of category y_c, and a certain dimension of $\boldsymbol{x}$ ($\boldsymbol{p}$) encodes a specific salient attribute. This novel prototype view has enabled researchers to develop many computational models for categorization, including the famous k-Nearest Neighbors (k-NN) and Nearest Centroids (Fix and Hodges, 1952; Cover and Hart, 1967). These prototype models mainly differ in how the prototypes are derived. For instance, the k-NN algorithm treats the k-nearest neighbors of the data samples as prototypes, while the Nearest Centroids algorithm considers the centroid, or average, of each category as the prototype.

In general, prototype theory is well-suited to explain the learning of many visual categories with a strong family-resemblance structure. However, **the prototype theory lacks the notion of scope**, making it less tolerant to intra-class variance.

From a statistical perspective, the use of prototype and scope to describe category is essentially to capture the form or structure of the data distribution $p(x|y)$, *i.e.*, how the data samples of a certain category look like. Hence the computational model of visual concept based categorization is a *generative classifier*, which estimates the conditional probability of a label given an input, and then uses Bayes' rule to assign the most likely label (Mackowiak et al., 2021):

$$p(y|\boldsymbol{x})=\frac{p(y)p(\boldsymbol{x}|y)}{\sum_{c\in\mathcal{Y}}p(c)p(\boldsymbol{x}|c)}. \tag{2}$$

Unlike *discriminative classifiers*, which directly map inputs to labels without explicit modeling of data distribution, generative classifiers are more difficult to train, because they have to model more aspects of the data (Liang et al., 2022a). This partially explains the challenges of constructing visual knowledge.

2.2.2 Visual relation

The term "visual relation" in visual knowledge theory denotes the connections and interactions that prevail among visual concepts, which are pivotal in navigating the complex landscape of visual cognition. Humans harbor an extensive repository of knowledge pertaining to the attributes of visual objects, a repository that transcends the mere intrinsic properties of these objects – such as color, shape, and texture – to include the relational properties interlinking them. These encompass their relative positioning, semantic dependencies, and affordances, which collectively constitute the relational properties, or visual relations. These relational properties, or visual relations, are amenable to categorization into distinct classes, each illuminating different facets of visual cognition:

Unedited

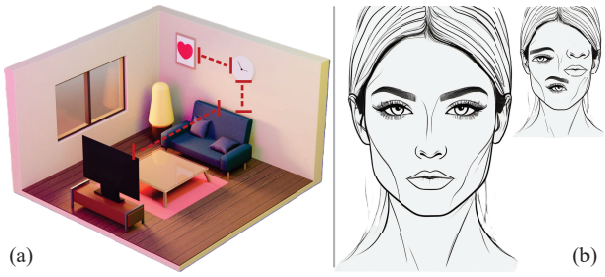

Fig. 3 Illustration of geometric relations.

Geometric relations: These relations delineate how objects or concepts are interconnected based on their spatial configurations and geometric constructs, such as their relative position, direction, distance, intersection, alignment, parallelism, perpendicularity, and so on (see Fig. 3(a)). Such relations facilitate an understanding of the structure and organization of objects within the environment, and unveil the inherent harmony and order of nature and art. For example, the kernels in an apple are located at its center. Similarly, our knowledge about the human face is not only the occurrence of eyes, nose, mouth, and ears, but also the precise spatial arrangement of those key facial elements (Fig. 3(b)).

Relation	Illustration	Interpretation
Before (A before B) After (B after A)	A B	A occurs completely before B B occurs completely after A
During (A during B) Contain (B contains A)	A B	A occurs completely within B, with B starting before A and ending after A B completely contains A, starting before A and ending after A
Overlap (A overlaps B) Overlapped by (B overlapped by A)	A B	A starts before B and ends after the start of B but before the end of B B starts after the start of A and ends after A has ended
Meet (A meets B) Met by (B met by A)	A B	A ends exactly when B begins B begins exactly when A ends
Start (A starts B) Started by (B started by A)	A B	A starts at the same time as B but ends before B B starts at the same time as A but ends after A
Finishes (A finishes B) Finished by (B finished by A)	A B	A starts and ends at the same time as B A starts after B has started and ends at the same time B ends
Equals (A equals B)	A B	B starts before A and ends at the same time as A

Fig. 4 Illustration of 13 base temporal relations defined in Allen's interval algebra (Allen, 1983).

Temporal relations: These relations, while not always directly visual, enrich visual knowledge by marking the sequence or timing of events and transformations within a visual scene over time. For example, temporal relations can describe the progression of actions, such as "before," "after," and "during," which are instrumental in comprehending the dynamics of environments and activities. Fig. 4 shows the 13 base temporal relations defined by Allen (1983).

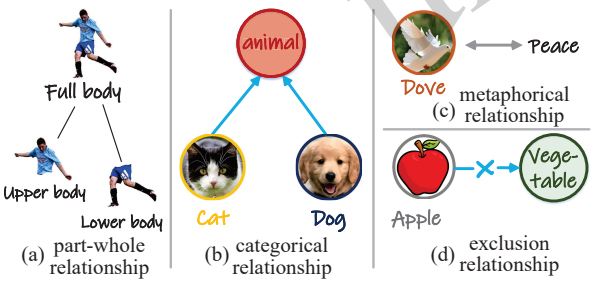

Fig. 5 Illustration of semantic relations.

Semantic relations: These relations specify the connections between objects or concepts grounded in their meanings or significances, hence enhancing our grasp of meanings, part-whole relationships, similarities, differences, inclusion-exclusion criteria, and semantic dependencies within visual information. For example, the part-whole relationships help us dissect visual concepts into their constituent parts or components, such as an apple can be broken down into its kernels, flesh, rind, and pedicel; human body can be broken down into different parts (Fig. 5(a)). Each of these sub-concepts maintains its own semantic connections both among themselves and with the overarching concept. In a similar vein, the categorical relationships describe the fact that visual concepts can be categorized with others that fall under the same category or superordinate concept; these concepts have semantic relations of similarity and difference with each other and with the category. For instance, dogs and cats are different types of animals (Fig. 5(b)). Moreover, semantic relationships involve more abstract associations, like the metaphorical relationship between "dove" and "peace" (Fig. 5(c)). Furthermore, semantic relationships enable the inclusion or exclusion of visual concepts based on specific criteria or rules, such as whether they belong to a certain domain or context. For example, apples and oranges are classified as fruits, not vegetables, based on certain distinguishing factors (Fig. 5(d)).

Fig. 6 Illustration of functional relations.

Functional relations: These relations explicate the interactions among objects based on their physical properties or affordances, hence facilitating our understanding of purpose, utility, effect, cause, and action-relevant structures. For example, a knife can cut bread, a chair can support a human, a pen can write on a paper, and so on. Functional relations establish the causal links between behavior and its environmental antecedents (stimuli) and consequences (reinforcers or punishers). For example, if a child learns that pressing a button produces a sound (antecedent), he or she may press the button more often (behavior) to hear the sound more frequently (consequence). Functional relations are foundational for reasoning and problem-solving, as they allow human to infer new facts or actions from existing facts or actions. By identifying the functional relations between problem behaviors and their environmental variables, one can design interventions that either change the antecedents or consequences of problem behaviors or teach alternative behaviors that serve the same function. For example, we can use functional relations to infer that if a knife can cut bread,

unedited

then it can also cut cheese. We can also use functional relations to infer that if we want to crush stones into pieces, we may need a hammer (Fig. 6). Functional relations also help in generating explanations or justifications for facts or actions. For example, functional relations can be used to explain why a knife is required to cut bread or why people sit on chairs.

Fig. 7 Illustration of causal relations.

Causal relations: These relations identify cause-and-effect connections between visual elements, which are essential for interpreting how and why changes occur in the visual context, such as understanding that rain makes the streets wet (Fig. 7). Causal relations also enable predictive reasoning regarding the outcomes of actions and events within a visual context.

Modeling these visual relations is fundamental to the fabric of the visual knowledge theory, as it enables AI systems to process and decipher visual information in a structured and meaningful manner. By categorizing and analyzing these relations, researchers can develop more sophisticated artificial models for visual perception, enhancing our ability to replicate human-like understanding and reasoning in machines.

2.2.3 Visual operation

The term "visual operation" in the visual knowledge theory denotes transformations over visual concepts or objects in space or time, such as composition, decomposition, replacement, combination, deformation, motion, comparison, destruction, restoration, and prediction. Visual concepts are the key elements of visual knowledge, enabling us to recognize, categorize, and name the entities we observe in our environment. Furthermore, visual relations enhance our understanding of the interconnectedness and functionalities of these entities. Yet, as illuminated by cognitive studies from scholars such as (Margolis and Laurence, 1999; Carey, 2000; Nersessian, 2010; Avargues-Weber et al., 2015; Thagard, 2013), visual concepts are subject to manipulation through cognitive processes that, for example, transform them in space or time, alter their components or characteristics, and facilitate various operations over these concepts or objects. These operations are instrumental in augmenting our capacity to comprehend the world, fostering innovation, and executing intricate tasks. They embody the dynamic aspect of visual knowledge, showcasing how static images or scenes can be reimagined or restructured through cognitive engagement:

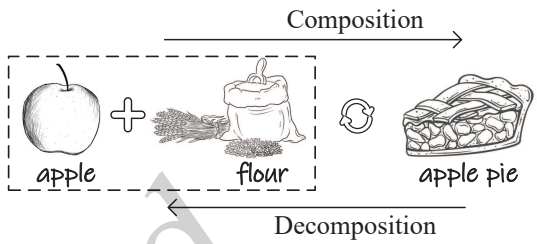

Fig. 8 Illustration of composition and decomposition operations.

Composition and decomposition: Composition involves assembling multiple visual elements to form a new object or concept, whereas decomposition refers to breaking down an object into its constituent elements. These operations are crucial for understanding complex systems and structures by analyzing their parts and how they fit together. Moreover, they are essential for the generation of innovative and creative concepts or objects. For example, through a thoughtful arrangement of an apple with other objects (*e.g.*, flour), new inventions be invented (*e.g.*, an apple pie; Fig. 8).

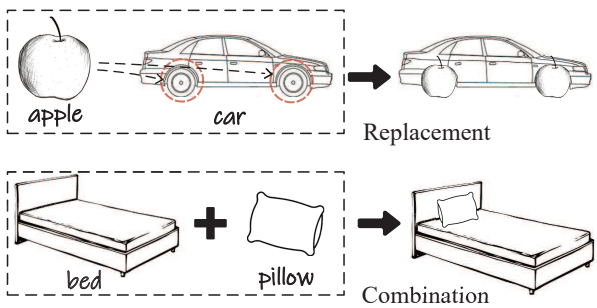

Fig. 9 Illustration of replacement and combination operations.

Replacement and combination: Replacement entails the substitution of one visual element by another, whereas combination pertains to merging

unedited

distinct elements to forge a new entity. These operations are fundamental to creative thinking and problem-solving, enabling the exploration of alternative configurations and solutions. They also enhance our understanding of the functionality of objects by allowing for imaginative scenarios, such as replacing the wheels of a car with apples (Fig. 9).

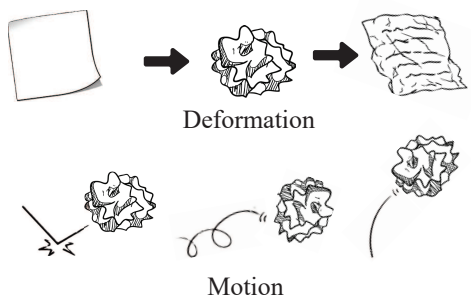

Fig. 10 Illustration of deformation and motion operations.

Deformation and motion: Deformation refers to altering the shape or structure of an object, whereas motion involves changing its position over time. Understanding these operations is vital for comprehending both the intrinsic and extrinsic properties of objects, interpreting various physical and biological processes, as well as for crafting animations and simulations that replicate real-world phenomena. Examples include manipulating an apple by scaling, rotating, or translating it in space, or modifying the motion of a falling apple by speeding up, slowing down, reversing, looping, or interpolating its moving trajectory (Fig. 10).

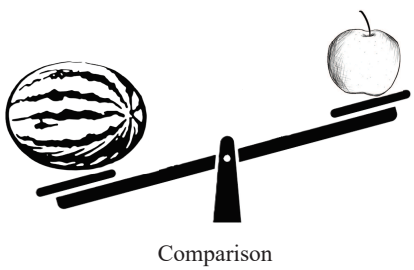

Fig. 11 Illustration of comparison operation.

Comparison: This operation entails evaluating similarities and differences between visual elements, aiding in classification, and decision-making processes. For example, we can compare an apple with other apples or objects in terms of size, weight, *etc* (Fig. 11). Comparison is essential for discerning patterns, making judgments, and learning from visual experiences.

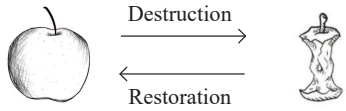

Fig. 12 Illustration of destruction and restoration operations.

Destruction and restoration: Destruction involves the removal or breakdown of visual elements, whereas restoration focuses on repairing or returning them to their original state (Fig. 12). These operations can be applied in various contexts, from understanding natural disasters and their aftermath to conservation efforts in art and historical preservation.

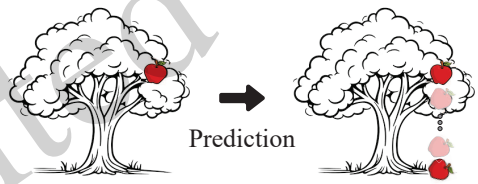

Fig. 13 Illustration of the prediction operation.

Prediction: This operation involves projecting future states or changes of visual elements based on current or past information (Fig. 13). This operation is crucial for planning, forecasting, and anticipating outcomes of actions and events.

Through these operations, the visual knowledge theory provides a framework for comprehending how visual information can be dynamically manipulated and utilized. These operations underscore the versatility and power of visual knowledge, illustrating its pivotal role in enhancing our ability to interact with, modify, and make predictions about the visual world, as well as its vast potential for application across various domains.

2.2.4 Visual reasoning

The term "visual reasoning" in the visual knowledge theory refers to the process of applying the knowledge gained from visual concepts, relations, and operations to interpret visual data, solve problems, and make informed decisions (Fig. 14). This intricate process typically entails a series of methodical operations on visual concepts and relations, aimed at

unedited

Fig. 14 Two examples for visual reasoning.

deriving valid and sound conclusions from what they observe visually and already know (common sense and knowledge).

In short, visual concept (Sec. 2.2.1) is about the identification and categorization of visual items; visual relation (Sec. 2.2.2) is about understanding the connections or associations between these items; visual operation (Sec. 2.2.3) is about the processes applied to manipulate or analyze visual items; and visual reasoning is about the process (Sec. 2.2.4) that uses visual concepts, relations, and operations to solve problems, make decisions, or derive sound conclusions from visual information.

3 Visual knowledge in the pre big model era: retrospect

In this section, we provide an overview of recent research relevant to visual knowledge, along the four key components of visual knowledge outlined in Sec. 2, namely visual concept, visual relation, visual operation, and visual reasoning.

3.1 Visual knowledge: visual concept

The idea of representing visual concepts by *prototype* and *scope* (Sec. 2.2.1) has been explored in a few fundamental computer vision tasks, namely image classification and segmentation. For instance, prototype-based networks (Snell et al., 2017), non-parametric neural classifiers (Zhou et al., 2022b), and nearest centroid-based neural classifiers (Wang et al., 2023) were developed, where each class is represented by one or a few prototypes, and new observations are classified by their proximity to the class prototypes. Despite their impressive performance in few-shot and general settings, these approaches fail to capture the scope of each class or prototype. For a more comprehensive modeling of the underlying data distribution, deep generative classifiers (Liang et al., 2022a) were devised to estimate the data density of each visual concept or class as a Gaussian Mixture Model (GMM), where the prototypes and scopes are the estimated GMM's parameters (*i.e.*, the mean vectors and covariance matrices). Its remarkable results on both closed-set and open-world scenarios evidence the power of prototype and scope based visual concept representation.

3.2 Visual knowledge: visual relation

As discussed in Sec. 2.2.2, visual concepts can be related to each other in different ways, resulting in various types of visual relations, such as geometric, temporal, semantic, functional, and causal relations.

Geometric relations describe how objects are arranged and transformed in space, including their position, orientation, size and shape. Capsule network (Sabour et al., 2017) is a landmark effort for modeling the geometric relations among visual elements. Basically, a capsule is a collection of neurons whose activity vector represents the probability and pose of a visual concept or object. A pose is a set of parameters describing the spatial relation and transformation of a visual concept, such as its location, rotation, scale and reflection. Though theoretical impressive, capsule networks are less practical for real-world applications, suggesting the great challenge of visual geometric relation modeling.

Semantic relations specify how objects are related to each other in terms of their meanings. A set of recent efforts are devoted to explore semantic relations in the context of visual understanding (Li et al., 2022, 2024) and human parsing (Wang et al., 2019, 2020). For instance, Li et al. (2022) proposed a neural parser that can generate structured, pixel-wise descriptions of visual observations in terms of a semantic concept hierarchy. This structured visual parser make an explicit use of the composition and decomposition dependencies among semantic concepts as additional regularization terms during network training. For example, an observation that is likely to be a cat should have a low probability of being any vehicle subcategory. However, it is important to note that the semantic relations encoded in the class hierarchy are predefined, rather than automatically learned. This suggests learning visual semantic relations is a challenging problem that requires much further research.

Temporal relations explicate the sequential or chronological order of events and actions as they oc-

unedited

cur over time within visual data. Temporal relations are mainly studied in the fields of action recognition and video object detection. For instance, in the field of action recognition, Something-Something dataset (Goyal et al., 2017) requires fine-grained motion distinctions and temporal modeling to distinguish interactions like *picking something up* and *putting something down*, providing a good testbed for temporal relation understanding. Video object detection is also a classic computer vision task with the target of classifying, segmenting, and tracking object instances in video sequences (Russakovsky et al., 2015; Zhou et al., 2023).

Functional relations refer to the actions that objects enable or support. Various tasks in computer vision study the functional relations of visual concepts, such as human-object interaction (HOI) detection, and affordance estimation (a.k.a functional recognition). HOI detection (Gupta et al., 2009; Zhou et al., 2022a; Li et al., 2023b) aims to locate and identify the relationships between humans and objects in visual scenes, like <*girl*, *eat*, *apple*>, while affordance estimation (Stark and Bowyer, 1991; Li et al., 2023) is to predict typical action-object affordances from visual information, such as *eatable*, *openable*.

Another related task to visual relation understanding is scene graph generation (SGG) (Johnson et al., 2015; Krishna et al., 2017), which is to generate a visually-grounded graph as an explicit structural description of a visual scene. The nodes in a scene graph represent the objects and the edges represent the relationships between objects (including spatial, part-whole, and interaction relationships). Each relationship between two objects is denoted as a triplet of <`subject`, `PREDICATE`, `object`>, *i.e.*, <`boy`, `RIDE`, `car`>, <`car`, `HAS`, `wheel`>, <`car`, `NEAR`, `building`>. Although SGG, to some extent, encompasses the three types of visual relations, namely geometric, semantic, and functional relations, in one single task, the covered relations are still sparse, for example, the functional relations are typically human-centric actions, less considering object-centric affordances. Moreover, SGG requires costly human efforts for visual relation annotation.

Causal relations delineate how events, actions, or objects within a visual context can directly influence or result in one another. Recent years witnessed unprecedented achievements of the deep learning technique across various domains, which, however, relies heavily upon fitting the data distributions. It is apparent that such technique tends to only learn correlation-based patterns (statistical dependencies) rather than the essential causal relationships from data; thus it easily collapses into data bias and has limited generalization ability (Hendrycks and Dietterich, 2019). In response to this challenge, a set of recent efforts have shifted toward uncovering the causality embedded in visual data. Basically, these efforts investigate causal reasoning (Pearl, 2009; Schölkopf et al., 2021) within the deep learning framework to extract causal representations from visual data. Leveraging causality-guided visual representations, they achieved notable performance in tasks such as visual recognition (Yue et al., 2021) and visual question answering (VQA) (Wang et al., 2020). They also showed that they are capable of automatically discovering causal dependencies among environmental and object variables from videos (Li et al., 2020), and improving the interpretability (Shi et al., 2022) and out-of-distributionÂăgeneralization ability (Christiansen et al., 2022) of deep learning models.

3.3 Visual knowledge: visual operation

As discussed in Sec. 2.2.3, visual operation refers to the manipulation of over visual concepts or objects in space or time. A closely related research field is customized visual content generation, aiming at generating creative contents for a target novel concept guided by textual descriptions. The textual descriptions serve as a feasible and flexible tool for specifying editing intensions, allowing for diverse visual operations such as replacement and combination. The task of generating realistic images from natural language descriptions – text-to-image synthesis – has been a research focus for years. Various deep generative models have been established for this task, such as GANs (Reed et al., 2016), VAEs (Ramesh et al., 2021), and Autoregressive Models (Ramesh et al., 2022). Recently, diffusion models have demonstrated a remarkable ability in generating text-aligned images with high fidelity (Rombach et al., 2022; Saharia et al., 2022). However, they encounter difficulties in performing customized generation for novel concepts, such as a specific animal or object which only appear in single testing image. Various customized visual content generation approaches have

unedited

been developed to address this challenge (Gal et al., 2023; Ruiz et al., 2023). They lift pre-trained text-to-image synthesis models to synthesize novel scenes of target concepts (typically represented by one or a few user-provided reference images) under natural language instruction. They have proven capable of generating not only creative static images (Gal et al., 2023; Ruiz et al., 2023) but also temporally coherent videos (Xing et al., 2024) that adhere to the guidance intentions. Though impressive, they may struggle for complicated visual operations (like decomposition, destruction, and restoration).

Novel view synthesis, *i.e.*, synthesizing new images of the same object or scene from arbitrary viewpoints given single or multiple inputs of the object/scene, is relevant to two visual operations: deformation and motion. Building on the concept of novel view synthesis, it becomes imperative to discuss the intricacies involved in simulating deformation and motion for objects within these synthesized viewpoints. This necessitates a deep understanding of the spatial and temporal aspects of objects, allowing for the generation of images that not only look realistic but also behave in ways that are consistent with the physical world. Techniques such as 3D modeling and neural radiance fields (NeRFs) (Mildenhall et al., 2020) have been instrumental in advancing this area. NeRFs, in particular, have shown great promise in creating detailed and continuous volumetric scenes, enabling smooth transitions and realistic deformations across different views (Pumarola et al., 2021). However, they require significant computational resources for both training and inference. Recently, 3D Gaussian splatting (Kerbl et al., 2023; Chen and Wang, 2024), which represents a 3D scene with millions of 3D Gaussians, has demonstrated a remarkable ability in real-time rendering. By introducing additional spatio-temporal modules (Yang et al., 2024a) or Gaussian properties (Luiten et al., 2023), they can model the dynamics and deformation in a given scene.

Regarding prediction, the visual operation of forecasting future states, actions, or events from visual data, has garnered significant interest, with numerous pertinent tasks in computer vision. Some representative tasks are:

- Human trajectory prediction: estimate the future paths of people in various settings, such as pedestrians on sidewalks or shoppers in malls, considering social behaviors and surroundings (Alahi et al., 2016).
- Future frame prediction: generate future frames of a video sequence to accurately depict the continuation of the observed scene (Mathieu et al., 2016).
- Action prediction: anticipate the future actions of subjects in a video, such as predicting an athlete's next move in sports or a driver's behavior in traffic (Ryoo, 2011).
- Physical interaction forecasting: predict the outcome of physical interactions between objects, such as forecasting the motion of objects after a collision (Watters et al., 2017).
- Accident anticipation: identify potentially hazardous situations before they occur, such as anticipating vehicle collisions or industrial accidents from a surveillance footage (Suzuki et al., 2018).

3.4 Visual knowledge: visual reasoning

Visual reasoning is the process of applying visual concept, visual relation, and visual operation, which are commonalities among various tasks, to draw valid and sound conclusions from premises or evidence (Pan, 1996). We can use functional relations to infer like "If A can cut B, then B is soft" or "If A can support B, then A is stable" (Sec. 2.2.3). Reasoning is a fundamental cognitive process that enables humans to make sense of the world. Human reasoning is a complex and multifaceted phenomenon that can be categorized into different types, such as deductive, inductive, abductive, and analogical, depending on the nature and strength of the arguments involved. Machine reasoning is a field of AI that complements the field of machine learning by aiming to perform automated reasoning. This is done by way of uniting known (yet possibly incomplete) information with background knowledge and making inferences regarding unknown or uncertain information, with the aid of efforts at many different disciplines, such as cognitive neuroscience, psychology, linguistics, and logic (Bottou, 2014).

Early explorations of automated reasoning systems primarily adopted two approaches: connectionism and symbolism. In the 1940s, McCulloch and Pitts proposed the first simplified neuronal model (McCulloch and Pitts, 1943), establishing the foundation for research in neural networks and connectionism. From the perspective of connectionism,

unedited

reasoning is the result or derivation of multiple, interconnected, simple processing devices, one major example being neural networks. The main motivation behind connectionism comes from cognitive neuroscience, since the human neural circuitry is clearly capable of storing and retrieving knowledge organized in short-and long-term memory, by continuously analyzing and processing new, complex information, and reasoning upon it. While connectionist models, particularly neural networks, are adept at capturing statistical patterns from data, they are bounded to computational resource and data availability at that time. Hence achieving the nuanced reasoning for human-like inferences remains intricate. This challenge prompted the exploration of symbolism, which has its roots in the study of logic and philosophy and became the dominate approach to Good-Old-Fashioned AI from the middle 1950s to the late 1980s. Basically, symbolic approaches posit that symbols representing worldly objects and concepts form the foundational building blocks of human intelligence. They see reasoning as the capability of deriving additional information from that already encoded in a collection of given symbols, by performing elaboration and manipulation on the given structured symbolic representations. They typically conduct reasoning by applying a series of restrict rules and formal logic operations to manipulate the discrete symbols in a rigorous and precise way. The rules defined how symbols could be manipulated to draw conclusions or make decisions. For example, a symbolic rule might deduce that "Socrates is mortal" from the premises that "All men are mortal" and "Socrates is a man." Consequently, symbolic approaches are powerful for problems with well-defined rules and discrete values, offering transparent explanations for their reasoning processes and allowing to check the validity or satisfiability of their logical steps. However, they are intolerant of ambiguous and noisy data, making them less suitable for many real-world applications due to the difficulty in defining the hard rules and high degree of uncertainty involved. Probabilistic reasoning techniques, such as Bayesian networks, Markov decision processes, and stochastic models, use the probability theory to represent and manipulate uncertainty, and can provide probabilistic estimates for the outcomes of reasoning. While these techniques can empower symbolic methods with the ability to deal with uncertainty, the intrinsic limitations of symbolic approaches, namely the lack of true learning and reliance of hand-crafted rules – remain significant obstacles to their application in practical contexts.

Recently, the emergence of large-scale datasets coupled with significant advancements in computational resources has sparked a resurgence in connectionism, particularly rejuvenating interest in neural network algorithms that are decades old. Yet, despite their prowess in pattern recognition and predictive modeling, DNNs struggle with reasoning tasks that necessitate explicit manipulation of symbols. Specifically, while DNNs excel at learning subsymbolics (*i.e.*, continuous embedding vectors), their architecture is not inherently suited for discrete symbolic operations that reasoning often entails. Moreover, DNNs typically learn from data in an inductive manner, which contrasts with the deductive procedure prevalent in reasoning, where logical deductions are drawn from explicit, pre-established rules and knowledge bases. It is not straightforward to integrate domain-specific knowledge into DNNs for explicit reasoning. In addition, the decision-making process of DNNs is often obscured, making it challenging to comprehend how particular conclusions are reached. This opacity is particularly problematic in decision-critical applications such as autonomous driving. On the other hand, symbolic approaches, though far less trainable, are excellent at principled judgements (such as deductive reasoning), and exhibit inherently high explainability (as they operates on clear, logical principles that can be easily traced and understood). In light of the challenges faced by DNNs in explicit reasoning and considering the complementary nature of connectionist and symbolic methodologies, a pioneering research domain, termed neuro-symbolic computing (NeSy), has gained prominence (Garcez et al., 2019). NeSy essentially pursues the principled integration of the two foundational paradigms in AI, providing a new framework of more powerful, transparent, and robust reasoning (Wang and Yang, 2022).

Traditionally, tasks related to visual reasoning are typically visual question-answering (VQA) and visual semantic parsing. VQA, *i.e.*, answering questions based on visual content, requires comprehensive understanding and reasoning over both the visual and linguistic modalities. Andreas et al. (2016) introduced a NeSy based VQA system that inter-

unedited

prets questions as executable programs composed of learnable neural modules that can be directly applied to images. A module is typically implemented by the neural attention operation and corresponds to a certain atomic reasoning step, such as recognizing objects, classifying colors, *etc*. This pioneering work stimulated many subsequent studies that explore NeSy for approaching VQA (Yi et al., 2018; Vedantam et al., 2019; Amizadeh et al., 2020). Visual semantic parsing seeks for a holistic explanation of visual observation in terms of a class hierarchy. The class hierarchy, pre-given as a knowledge base, encapsulates the symbolic relations among semantic concepts. Li et al. (2023a) devise a powerful NeSy based visual semantic parser through end-to-end embedding symbolic logic into both the network's training and inference stages. Some other relevant tasks include visual abductive reasoning (Liang et al., 2022b) and visual commonsense reasoning (Zellers et al., 2019).

More recently, benefiting from the impressive emergent capabilities of large language models (LLM), some efforts explore LLMs for solving sophisticated visual reasoning tasks. VisProg (Gupta and Kembhavi, 2023) represents a pioneering effort in this domain; it uses LLMs to decompose visual reasoning tasks (such as "Is it true that the two images contain a total of six people and two boats?") into a series of manageable subtasks (such as text parsing, object detection, and counting) and solving them step-by-step. Later, HuggingGPT (Shen et al., 2023) leverages LLMs to manage AI models available on the web to solve complicated reasoning tasks. DoraemonGPT (Yang et al., 2024b) advances this research trend towards solving real-world tasks that involve dynamic observations. It equips LLMs with a symbolic memory for gathering and storing task-relevant information from the dynamic observation, a rich set of extra knowledge sources (*e.g.*, AI tools, search engines, text books, knowledge databases) for reference, and a Monte Carlo Tree Search based planner for efficiently probing the huge solution space.

3.5 Discussion

So far, we have reviewed prior key contributions relevant to the four foundational aspects of visual knowledge. From such review, some key insights can be derived as follows:

First, visual knowledge is closely linked to the two fundamental AI paradigms (namely connectionism and symbolism), several research domains (including computer vision, graphics, machine learning, and logic), a set of fundamental and challenging tasks (including visual recognition, affordance estimation, text-to-image synthesis, novel view synthesis, and future forecasting), and various advanced techniques (including capsule networks, neuro-symbolic computing, and large language models). This underscores the critical significance of visual knowledge, and the necessity of interdisciplinary collaborations to achieve this grand goal.

Second, while our community has indeed achieved progress in certain areas related to visual knowledge, numerous core issues remain challenging and underexplored, such as prototype-and scope-based visual concept, causal relation, complex visual operations (*e.g.*, decomposition, destruction, and restoration), and visual reasoning. This highlights the difficulties in creating visual knowledge. This also unveils a primary motivation for proposing visual knowledge – the lack of a principled framework that offers a unified perspective on different aspects of visual intelligence.

Third, although LLMs have demonstrated remarkable capabilities in solving intricate problems, they also exacerbate the "black box" issue, an innate character of neural network algorithms. With their billions or even trillions of parameters, LLMs pose an insurmountable challenge for anyone attempting to dissect their internal workings. Moreover, LLMs are still yet to perform logical reasoning in the way humans do. They frequently produce plausible-sounding answers that, upon closer examination, reveal a lack of genuine comprehension. Compounding this issue is the opaque nature of LLMs, which obstructs the identification and correction of errors within the reasoning process. In the subsequent section, we will delve into the significance of investigating LLMs within the visual knowledge framework.

4 Visual knowledge in the big model era: prospect

In this section, we first explore how visual knowledge can power big models to bring the level of general intelligence closer to that of humans. We then investigate how big models can boost the development of visual knowledge, given the significant

unedited

challenges of establishing visual knowledge.

4.1 Empower big models with visual knowledge

The advent of large AI models has undeniably marked a new era in AI, providing unprecedented accuracy and fluency in tasks that were once considered insurmountable for machine. However, despite their astonishing successes, these powerful models are not without their limitations. Among the most critical challenges they face are issues related to transparency, reasoning, and catastrophic forgetting. In the forthcoming discussion, we will show that, despite these substantial obstacles, the integration of visual knowledge into the large models offers a promising avenue for advancement.

One of the most discussed limitations of large models is their lack of transparency. Transparency refers to the degree to which the internal workings and outputs of models can be understood and explained by humans. Due to the sheer volume of parameters, understanding how the big models arrive at a particular conclusion is extremely challenging. The lack of transparency hinders our ability to fully trust and verify the models' decisions, especially in critical applications such as healthcare diagnostics, and autonomous driving, and causes a lot of concerns regarding accountability, bias, fairness, debugging, *etc.* Although there are some network interpretation techniques based upon the analysis of reverse-engineer importance values or sensitivities of inputs, they only produce posteriori explanations for already-trained DNNs. They essentially *approximate* the behavior of DNN by modeling relationships between features and the outputs. Such *post-hoc* explanations are problematic and misleading, as they cannot explain what actually makes a DNN arrive at its decisions (Rudin et al., 2022). Yet, due to the inherent transparent nature of prototype-and scope-based visual concept, the integration of visual knowledge may endow big models with promising *ad-hoc* interpretability. A notable evidence is the groundbreaking study by Wang et al. (2023), which introduces Deep Nearest Centroids (DNC), a fully end-to-end, prototype-based neural classifier. Through representing visual concepts as a collection of automatically discovered prototypes (*i.e.*, class subcentroids), DNC mirrors the experience-/case-based reasoning process that humans are accustomed to and yields a powerful yet *ad-hoc* interpretable framework for large-scale visual recognition. This idea can be further explored for prototype-and scope-based visual concept modeling. By employing such inherently transparent visual concepts as foundational elements, it is naturally anticipated that visual knowledge will enhance the transparency of big models.

While large AI models excel at pattern recognition and generating human-like text or images, they may fail to grasp the underlying logic or truth of the content it produces; hence they may struggle with tasks requiring an understanding of causality, abstract concepts, or logical inference. For instance, big models may generate plausible-sounding but factually incorrect or nonsensical answers, known as "hallucination", reflecting a surface-level mimicry of human output, rather than a genuine comprehension (Ji et al., 2023). The challenge stems from the big models' reliance on statistical patterns or superficial features, rather than causation which may not capture the underlying causal relationships or logic. Although a few reasoning strategies such as Chain of Thoughts (Wei et al., 2022) and Tree of Thoughts (Yao et al., 2023) are recently developed for boosting large models' reasoning ability, they are still far from the true reasoning that typically involves managing complex operations over symbolic concepts, understanding causality, and applying abstract principles to novel situations. Yet, visual knowledge provides an explicit, powerful, and unified framework for the comprehensive modeling of visual conceptions, visual relations (including causality), visual operations, and visual reasoning. As a result, this may bring the reasoning of big models into a brand-new era where the reasoning is powered by both implicit knowledge from big models and explicit knowledge modeled by visual knowledge. In such wise, big models can perform complicated tasks like humans, *e.g.*, reasoning about complex and dynamic scenarios that involve multiple entities and complex relations, solving problems that require a series of methodical operations on visual concepts and relations, and applying learned knowledge to solve fundamentally different problems. Given the recent impressive progress in the integration of symbolic-knowledge based logic reasoning and data-driven neural subsymbolic learning (Li et al., 2023a), we firmly believe that combing big models' implicit knowledge and explicit visual knowledge in a form of multiple knowledge representation (Pan, 2020; Yang

unedited

et al., 2021) is a promising pathway forward.

Catastrophic forgetting refers to the tendency of DNNs to lose their previously learned knowledge when exposed to new data or tasks. The root of this issue lies in the way DNNs update their parameters; new learning can overwrite the weights and biases associated with old knowledge, leading to a rapid degradation of performance on tasks the model had previously mastered. Catastrophic forgetting is a fundamental challenge to the vision of creating AI systems that can learn and adapt over time in a manner analogous to human learning. At the heart of catastrophic forgetting lies the difficulty of knowledge trace within big AI models. Knowledge trace, or the ability to identify, follow, and understand the representation and processing of information within a model, is about knowing what the model knows and how it came to know it. In human learning, knowledge trace allows for the accumulation of experiences and the seamless integration of new information with existing knowledge. However, in large AI models, identifying the specific components responsible for particular pieces of knowledge is a daunting task due to the complex network architecture of massive interconnected parameters. Visual knowledge, with its deep root in cognitive psychology, offers large models with a form of knowledge representation that is explicit, structured, persistent, editable, and traceable. This allows to update knowledge outside the big models, enabling more targeted interventions to prevent catastrophic forgetting. Moreover, augmented with visual knowledge, the big models can create more durable and retrievable memory to enhance recall and understanding, like humans.

4.2 Boost visual knowledge with big models

Having highlighted the great importance of visual knowledge in enhancing big AI models, we will next explore the pivotal role that big models play in advancing visual knowledge.

First, big AI models will be an essential cornerstone of visual knowledge. Big models exhibit the unparalleled ability to discern meaningful patterns from large-scale data. Therefore, it is a nature choice to utilize the large-scale learning ability of big models to learn robust visual concepts, and model basic visual relations (such as temporal relations and geometric relations) and operations (such as composition, deformation, and motion).

Second, big AI models will serve as a knowledge source for visual knowledge. Trained on large volumes of text including scientific articles, Wikipedia, books, and other sources of information, large language models have been observed to learn not only contextualized text representations but also a significant body of world knowledge and commonsense knowledge (Safavi and Koutra, 2021). This suggests the great potential of big models as a knowledge base that could significantly enrich visual knowledge. For example, large language models can help to better capture semantic relationships between concepts, which are often not immediately apparent in visual data; the categorial relationship between “cat” and “animal” is more readily understood through text. However, unlocking this treasure trove of knowledge is a highly challenging task. The knowledge acquired by big models is deeply hidden within the network parameters, and it is not directly accessible for analysis and reuse. Moreover, the knowledge in big models is not purely factual or reliable; they are intertwined with errors, bias, noise, and trivial patterns. Therefore, advanced techniques (AlKhamissi et al., 2022) for knowledge analysis (identify and localize what knowledge has been acquired by language models), knowledge extraction (extract and represent the knowledge encoded in large models), and knowledge enhancement (validate and refine the extracted knowledge) should be utilized.

Third, big AI models will provide complementary knowledge for visual knowledge. Large language models model the world, as described by the text data. The knowledge acquired from the text data does not only enrich, but also complements visual knowledge. For example, some knowledge is hard to learn from visual data, such as human internal thoughts, motivations, and emotions, and commonsense knowledge like “Beijing is the caption of China”. Similarly, while a photograph of the Great Wall of China conveys its majesty and scale, it is through textual data that we learn about its historical significance, the reasons for its construction, and its role in Chinese culture. Cognitive studies also suggest visual memory and linguistic memory are not independent but interact with each other in complex ways. Visual memory can be influenced by linguistic information that provides meaning and context to visual stimuli. Linguistic memory can be influenced by visual information that provides imagery

unedited

and detail to verbal stimuli. As a result, complementing visual knowledge with the knowledge in big models will lead to a more holistic understanding of the world.

5 Conclusion

The last decade has witnessed breakthroughs in the field of AI, especially connectionist approaches to long standing challenges around computer vision, natural language processing, speech recognition, and autonomous systems. As the unique product of an era with a treasure trove of data from the internet and increasingly powerful computing resources, big AI models, which assemble the characteristics of both connectionism and scaling law, are swiftly embedding themselves into the fabric of human society and becoming indispensable for scientific discovery. While these developments are nothing short of revolutionary, and have fundamentally altered our way of life, there is a general agreement that this is just the beginning of an AI revolution. However, big AI models still exhibit deficiencies in, for example, transparency, accountability, and symbolic reasoning. Given the significant advantages of the comprehensive modeling of visual concepts, relations, operations, and reasoning, visual knowledge shows the promise of mitigating the shortcomings of existing AI techniques, unlocking the door of the next-generation AI. Starting from reviewing cognitive studies in visual memory and perception, this article introduces the origin and core concepts of visual knowledge. Subsequently, this article retrospects recent research efforts that are relevant to visual knowledge, along the dimensions of visual concept, visual relation, visual operation, and visual reasoning. Based on the analysis of the current research situation of visual knowledge, we point out the opportunities and challenges it faces in the era of big models, to pave the way research on the next-generation AI.

Contributors

Wenguan Wang, Yi Yang, and Yunhe Pan designed the research. Wenguan Wang drafted the manuscript. Yi Yang and Yunhe Pan revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

Data available within the article or its supplementary materials. The authors confirm that the data supporting the findings of this study are available within the article [and/or] its supplementary materials.

References

Alahi A, Goel K, Ramanathan V, et al., 2016. Social LSTM: human trajectory prediction in crowded spaces. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.961-971, https://doi.org/10.1109/CVPR.2016.110

AlKhamissi B, Li M, Celikyilmaz A, et al., 2022. A review on language models as knowledge bases. *CoRR*, abs/2204.06031. https://doi.org/10.48550/arXiv.2204.06031

Allen JF, 1983. Maintaining knowledge about temporal intervals. *Commun ACM*, 26(11):832-843. https://doi.org/10.1145/182.358434

Amizadeh S, Palangi H, Polozov A, et al., 2020. Neuro-symbolic visual reasoning: Disentangling "visual" from "reasoning". International Conference on Machine Learning, 119:279-290, http://proceedings.mlr.press/v119/amizadeh20a.html

Anderson JR, 2005. Cognitive psychology and its implications. Macmillan.

Andreas J, Rohrbach M, Darrell T, et al., 2016. Neural module networks. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.39-48, https://doi.org/10.1109/CVPR.2016.12

Avargues-Weber A, Giurfa M, Plotnik J, et al., 2015. The conceptual mind: New directions in the study of concepts. MIT Press.

Bigelow J, Poremba A, 2014. AchillesâĂŹ ear? inferior human short-term and recognition memory in the auditory modality. *PloS one*, 9(2):e89914. https://doi.org/10.1371/journal.pone.0089914

Bottou L, 2014. From machine learning to machine reasoning - an essay. *Mach Learn*, 94(2):133-149. https://doi.org/10.1007/s10994-013-5335-x

Brady TF, Konkle T, Alvarez GA, et al., 2008. Compression in visual short-term memory: Using statistical regularities to form more efficient memory representations. *Journal of Vision*, 8(6):199-199. https://doi.org/10.1167/8.6.199

Brady TF, Konkle T, Alvarez GA, 2011. A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, 11(5):4-4. https://doi.org/10.1167/11.5.4

Brown TB, Mann B, Ryder N, et al., 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems.

Carey S, 2000. The origin of concepts. *Journal of Cognition and Development*, 1(1):37-41.

Chen G, Wang W, 2024. A survey on 3d gaussian splatting. *CoRR*, abs/2401.03890. https://doi.org/10.48550/arXiv.2401.03890

unedited

Christiansen R, Pfister N, Jakobsen ME, et al., 2022. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614-6630. https://doi.org/10.1109/TPAMI.2021.3094760

Cover TM, Hart PE, 1967. Nearest neighbor pattern classification. *IEEE Trans Inf Theory*, 13(1):21-27. https://doi.org/10.1109/TIT.1967.1053964

Fix E, Hodges J, 1952. Discriminatory Analysis: Nonparametric Discrimination, Small Sample Performance. Air University, USAF School of Aviation Medecine, https://books.google.com.hk/books?id=bEsYnQEACAAJ

Gal R, Alaluf Y, Atzmon Y, et al., 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. International Conference on Learning Representations, https://openreview.net/pdf?id=NAQvF08TcyG

d'Avila Garcez AS, Gori M, Lamb LC, et al., 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6(4):611-632. https://collegepublications.co.uk/ifcolog/?00033

Goyal R, Kahou SE, Michalski V, et al., 2017. The "something something" video database for learning and evaluating visual common sense. International Conference on Computer Vision, p.5843-5851, https://doi.org/10.1109/ICCV.2017.622

Gupta A, Kembhavi A, Davis LS, 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775-1789. https://doi.org/10.1109/TPAMI.2009.83

Gupta T, Kembhavi A, 2023. Visual programming: Compositional visual reasoning without training. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.14953-14962, https://doi.org/10.1109/CVPR52729.2023.01436

Hendrycks D, Dietterich TG, 2019. Benchmarking neural network robustness to common corruptions and perturbations. International Conference on Learning Representations.

Ji Z, Lee N, Frieske R, et al., 2023. Survey of hallucination in natural language generation. *ACM Comput Surv*, 55(12):248:1-248:38. https://doi.org/10.1145/3571730

Johnson J, Krishna R, Stark M, et al., 2015. Image retrieval using scene graphs. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.3668-3678, https://doi.org/10.1109/CVPR.2015.7298990

Kerbl B, Kopanas G, Leimkühler T, et al., 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans Graph*, 42(4):139:1-139:14. https://doi.org/10.1145/3592433

Kirillov A, Mintun E, Ravi N, et al., 2023. Segment anything. International Conference on Computer Vision, p.3992-4003, https://doi.org/10.1109/ICCV51070.2023.00371

Kosslyn SM, Ball TM, Reiser BJ, 1978. Visual images preserve metric spatial information: evidence from studies of image scanning. *Journal of experimental psychology: Human perception and performance*, 4(1):47. https://doi.org/10.1037/0096-1523.4.1.47

Krishna R, Zhu Y, Groth O, et al., 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32-73. https://doi.org/10.1007/s11263-016-0981-7

Li L, Zhou T, Wang W, et al., 2022. Deep hierarchical semantic segmentation. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.1236-1247, https://doi.org/10.1109/CVPR52688.2022.00131

Li L, Wang W, Yi Y, 2023a. Logicseg: Parsing visual semantics with neural logic learning and reasoning. International Conference on Computer Vision, p.4099-4110, https://doi.org/10.1109/ICCV51070.2023.00381

Li L, Wei J, Wang W, et al., 2023b. Neural-logic human-object interaction detection. Advances in Neural Information Processing Systems.

Li L, Wang W, Zhou T, et al., 2024. Semantic hierarchy-aware segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2123-2138. https://doi.org/10.1109/TPAMI.2023.3332435

Li Y, Xu Y, Xu X, et al., 2023. Beyond object recognition: A new benchmark towards object concept learning. International Conference on Computer Vision, p.19972-19983, https://doi.org/10.1109/ICCV51070.2023.01833

Li Y, Torralba A, Anandkumar A, et al., 2020. Causal discovery in physical systems from videos. Advances in Neural Information Processing Systems.

Liang C, Wang W, Miao J, et al., 2022a. Gmmseg: Gaussian mixture based generative semantic segmentation models. Advances in Neural Information Processing Systems.

Liang C, Wang W, Zhou T, et al., 2022b. Visual abductive reasoning. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.15544-15554, https://doi.org/10.1109/CVPR52688.2022.01512

Luiten J, Kopanas G, Leibe B, et al., 2023. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *CoRR*, abs/2308.09713. https://doi.org/10.48550/arXiv.2308.09713

Mackowiak R, Ardizzone L, Köthe U, et al., 2021. Generative classifiers as a basis for trustworthy image classification. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.2971-2981, https://doi.org/10.1109/CVPR46437.2021.00299

Margolis EE, Laurence SE, 1999. Concepts: core readings. The MIT Press.

Mathieu M, Couprie C, LeCun Y, 2016. Deep multi-scale video prediction beyond mean square error. International Conference on Learning Representations.

McCulloch WS, Pitts W, 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115-133. https://doi.org/10.1007/BF02478259

Mildenhall B, Srinivasan PP, Tancik M, et al., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. European Conference on Computer Vision, 12346:405-421, https://doi.org/10.1007/978-3-030-58452-8_24

Milner D, Goodale M, 2006. The visual brain in action. Oup Oxford, https://doi.org/10.1093/acprof:oso/9780198524724.001.0001

unedited

Moyer RS, 1973. Comparing objects in memory: Evidence suggesting an internal psychophysics. *Perception & Psychophysics*, 13:180-184. https://doi.org/10.3758/BF03214124

Nersessian NJ, 2010. Creating scientific concepts. MIT press.

Pan Yh, 2019. On visual knowledge. *Front Inf Technol Electron Eng*, 20(8):1021-1025. https://doi.org/10.1631/FITEE.1910001

Pan Y, 1996. The synthesis reasoning. *Pattern Recognition and Artificial Intelligence (in Chinese)*, 9(3):8.

Pan Y, 2020. Multiple knowledge representation of artificial intelligence. *Engineering*, 6(3):216-217. https://doi.org/10.1016/j.eng.2019.12.011

Pearl J, 2009. Causality. Cambridge university press, https://doi.org/10.1017/CBO9780511803161

Pumarola A, Corona E, Pons-Moll G, et al., 2021. D-nerf: Neural radiance fields for dynamic scenes. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.10318-10327, https://doi.org/10.1109/CVPR46437.2021.01018

Ramesh A, Pavlov M, Goh G, et al., 2021. Zero-shot text-to-image generation. International Conference on Machine Learning, 139:8821-8831.

Ramesh A, Dhariwal P, Nichol A, et al., 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125. https://doi.org/10.48550/arXiv.2204.06125

Reed SE, Akata Z, Yan X, et al., 2016. Generative adversarial text to image synthesis. International Conference on Machine Learning, 48:1060-1069.

Rombach R, Blattmann A, Lorenz D, et al., 2022. High-resolution image synthesis with latent diffusion models. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.10674-10685, https://doi.org/10.1109/CVPR52688.2022.01042

Rosch E, Mervis CB, 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573-605. https://doi.org/10.1016/0010-0285(75)90024-9

Rudin C, Chen C, Chen Z, et al., 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1-85. https://doi.org/10.1214/21-SS133

Ruiz N, Li Y, Jampani V, et al., 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.22500-22510, https://doi.org/10.1109/CVPR52729.2023.02155

Russakovsky O, Deng J, Su H, et al., 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211-252. https://doi.org/10.1007/s11263-015-0816-y

Ryoo MS, 2011. Human activity prediction: Early recognition of ongoing activities from streaming videos. International Conference on Computer Vision, p.1036-1043, https://doi.org/10.1109/ICCV.2011.6126349

Sabour S, Frosst N, Hinton GE, 2017. Dynamic routing between capsules. Advances in Neural Information Processing Systems, p.3856-3866.

Safavi T, Koutra D, 2021. Relational world knowledge representation in contextual language models: A review. Conference on Empirical Methods in Natural Language Processing, p.1053-1067, https://doi.org/10.18653/v1/2021.emnlp-main.81

Saharia C, Chan W, Saxena S, et al., 2022. Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems.

Schölkopf B, Locatello F, Bauer S, et al., 2021. Toward causal representation learning. *Proc IEEE*, 109(5):612-634. https://doi.org/10.1109/JPROC.2021.3058954

Scolari M, Vogel EK, Awh E, 2008. Perceptual expertise enhances the resolution but not the number of representations in working memory. *Psychonomic bulletin & review*, 15(1):215-222. https://doi.org/10.3758/PBR.15.1.215

Shen Y, Song K, Tan X, et al., 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580. https://doi.org/10.48550/arXiv.2303.17580

Shepard RN, Feng C, 1972. A chronometric study of mental paper folding. *Cognitive psychology*, 3(2):228-243. https://doi.org/10.1016/0010-0285(72)90005-9

Shepard S, Metzler D, 1988. Mental rotation: effects of dimensionality of objects and type of task. *Journal of experimental psychology: Human perception and performance*, 14(1):3. https://doi.org/10.1037/0096-1523.14.1.3

Shi W, Huang G, Song S, et al., 2022. Temporal-spatial causal interpretations for vision-based reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10222-10235. https://doi.org/10.1109/TPAMI.2021.3133717

Snell J, Swersky K, Zemel RS, 2017. Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems, p.4077-4087.

Stark L, Bowyer KW, 1991. Achieving generalized object recognition through reasoning about association of function to structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1097-1104. https://doi.org/10.1109/34.99242

Suzuki T, Kataoka H, Aoki Y, et al., 2018. Anticipating traffic accidents with adaptive loss and large-scale incident DB. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.3521-3529, https://doi.org/10.1109/CVPR.2018.00371

Thagard P, 2013. Cognitive science, :597-608. https://doi.org/10.4324/9780203744857

Vedantam R, Desai K, Lee S, et al., 2019. Probabilistic neural symbolic models for interpretable visual question answering. International Conference on Machine Learning, 97:6428-6437.

Wang T, Huang J, Zhang H, et al., 2020. Visual commonsense R-CNN. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.10757-10767, https://doi.org/10.1109/CVPR42600.2020.01077

Wang W, Yang Y, 2022. Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing. *CoRR*, abs/2210.15889. https://doi.org/10.48550/arXiv.2210.15889

unedited

Wang W, Zhang Z, Qi S, et al., 2019. Learning compositional neural information fusion for human parsing. International Conference on Computer Vision, p.5702-5712, https://doi.org/10.1109/ICCV.2019.00580

Wang W, Zhu H, Dai J, et al., 2020. Hierarchical human parsing with typed part-relation reasoning. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.8926-8936, https://doi.org/10.1109/CVPR42600.2020.00895

Wang W, Han C, Zhou T, et al., 2023. Visual recognition with deep nearest centroids. International Conference on Learning Representations.

Watters N, Zoran D, Weber T, et al., 2017. Visual interaction networks: Learning a physics simulator from video. Advances in Neural Information Processing Systems, p.4539-4547.

Wei J, Wang X, Schuurmans D, et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems.

Xing J, Xia M, Liu Y, et al., 2024. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Trans Vis Comput Graph*, :1-15. https://doi.org/10.1109/TVCG.2024.3365804

Yang Y, Zhuang Y, Pan Y, 2021. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. *Frontiers of Information Technology & Electronic Engineering*, 22(12):1551-1558. https://doi.org/10.1631/FITEE.2100463

Yang Z, Yang H, Pan Z, et al., 2024a. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *International Conference on Learning Representations*, . https://doi.org/10.48550/arXiv.2310.10642

Yang Z, Chen G, Li X, et al., 2024b. Doraemongpt: Toward understanding dynamic scenes with large language models. *International Conference on Machine Learning*, . https://doi.org/10.48550/arXiv.2401.08392

Yao S, Yu D, Zhao J, et al., 2023. Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems.

Yi K, Wu J, Gan C, et al., 2018. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. Advances in Neural Information Processing Systems, p.1039-1050.

Yue Z, Wang T, Sun Q, et al., 2021. Counterfactual zero-shot and open-set visual recognition. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.15404-15414, https://doi.org/10.1109/CVPR46437.2021.01515

Zellers R, Bisk Y, Farhadi A, et al., 2019. From recognition to cognition: Visual commonsense reasoning. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.6720-6731, https://doi.org/10.1109/CVPR.2019.00688

Zhou T, Qi S, Wang W, et al., 2022a. Cascaded parsing of human-object interaction recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2827-2840. https://doi.org/10.1109/TPAMI.2021.3049156

Zhou T, Wang W, Konukoglu E, et al., 2022b. Rethinking semantic segmentation: A prototype view. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.2572-2583, https://doi.org/10.1109/CVPR52688.2022.00261

Zhou T, Porikli F, Crandall DJ, et al., 2023. A survey on deep learning technique for video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7099-7122. https://doi.org/10.1109/TPAMI.2022.3225573

unedited