Wenguan WANG, Yi YANG, Yunhe PAN, 2025. Visual knowledge in the big model era: retrospect and prospect. *Frontiers of Information Technology & Electronic Engineering*, 26(1):1-19. https://doi.org/10.1631/FITEE.2400250

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



**Key words:** Visual knowledge; Artificial intelligence; Foundation model; Deep learning

Corresponding author: Yi YANG
E-mail: yangyics@zju.edu.cn
ORCID: https://orcid.org/0000-0002-0512-880X

# Visual knowledge

- A new form of knowledge representation[1] that can encapsulate visual concepts and their relations in a succinct, comprehensive, and interpretable manner.

- A combination of four key components, namely ①**visual concept**, ②**visual relation**, ③**visual operation**, and ④**visual reasoning**.

- Playing a pivotal role in establishing machine intelligence — an indispensable component of human cognition and intelligence.

- Enabling AI systems to comprehensively describe, robustly recognize, and reason about visual items and solve tasks.

# Definition

## ① Visual concept

A category of **visual objects** that share some common features.

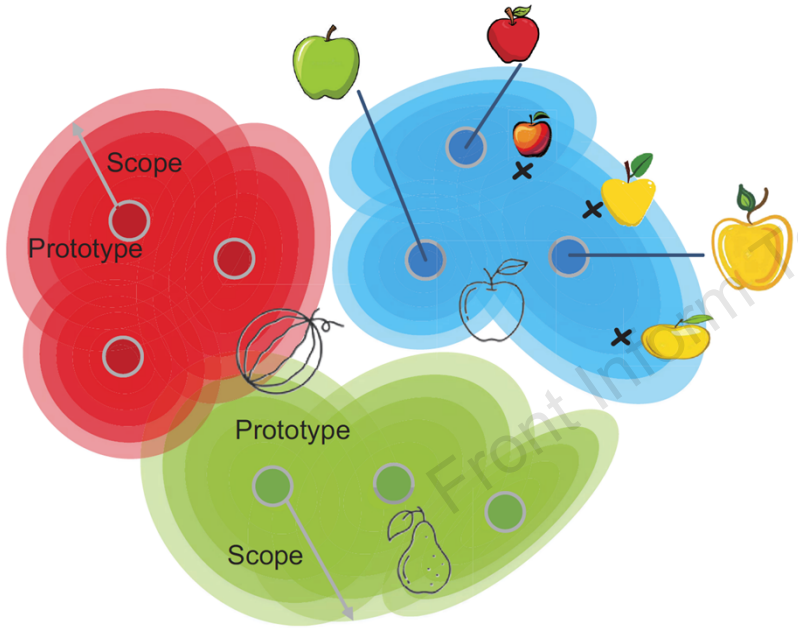
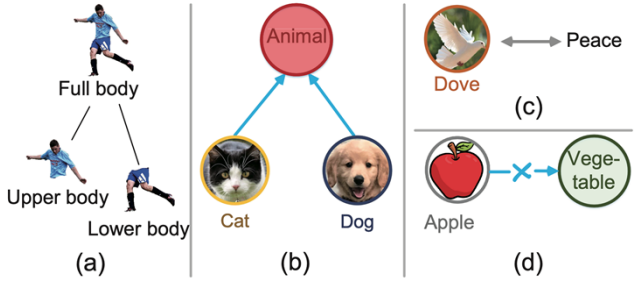



Illustration of prototype- and scope-based visual concept representation. Here we show three visual concepts, namely pear, apple, and watermelon.

## ② Visual relation

The connections and interactions that prevail among **visual concepts**, which are pivotal in navigating the complex landscape of visual cognition.


Geometric relations

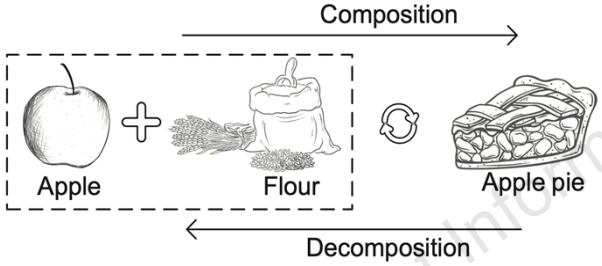



Semantic relations

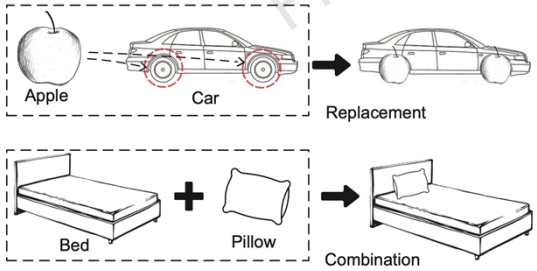
# Definition

## ③ Visual operation

Transformations over **visual concepts** or objects in space or time, such as composition, combination, deformation, decomposition, replacement, motion, comparison, destruction, restoration, and prediction.




Composition and decomposition




Replacement and combination

## ④ Visual reasoning

Applying the knowledge gained from **visual concepts, relations,** and **operations** to interpret visual data, solve problems, and make informed decisions.


Two examples for visual reasoning

# Visual knowledge in the pre-big-model era: retrospect

## ① Visual concept

*Explored in a few fundamental computer vision tasks.*

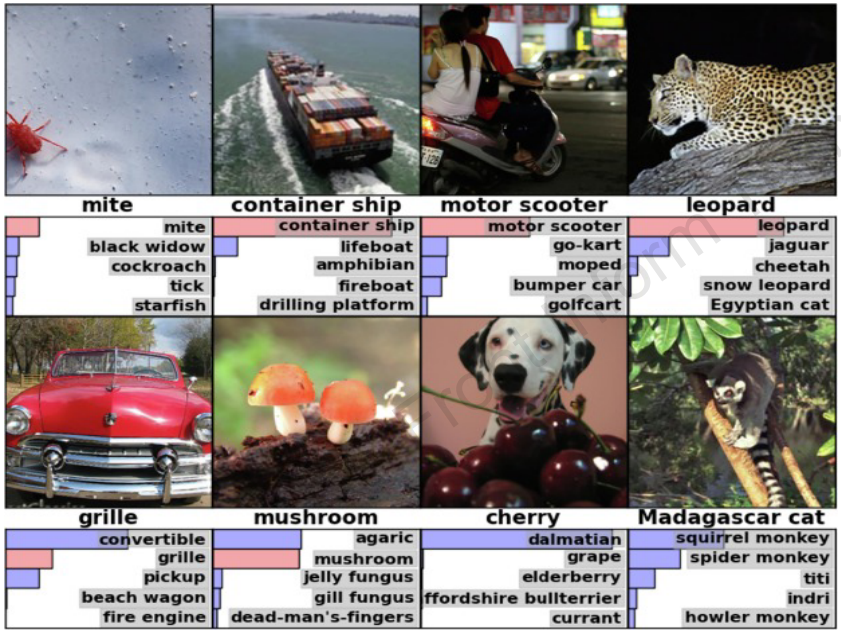



Image classification

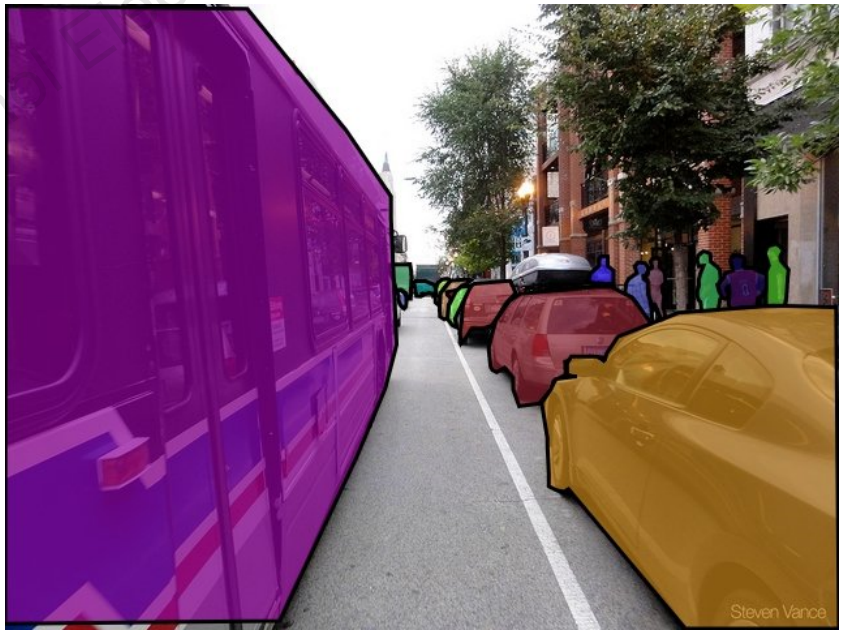



Image segmentation

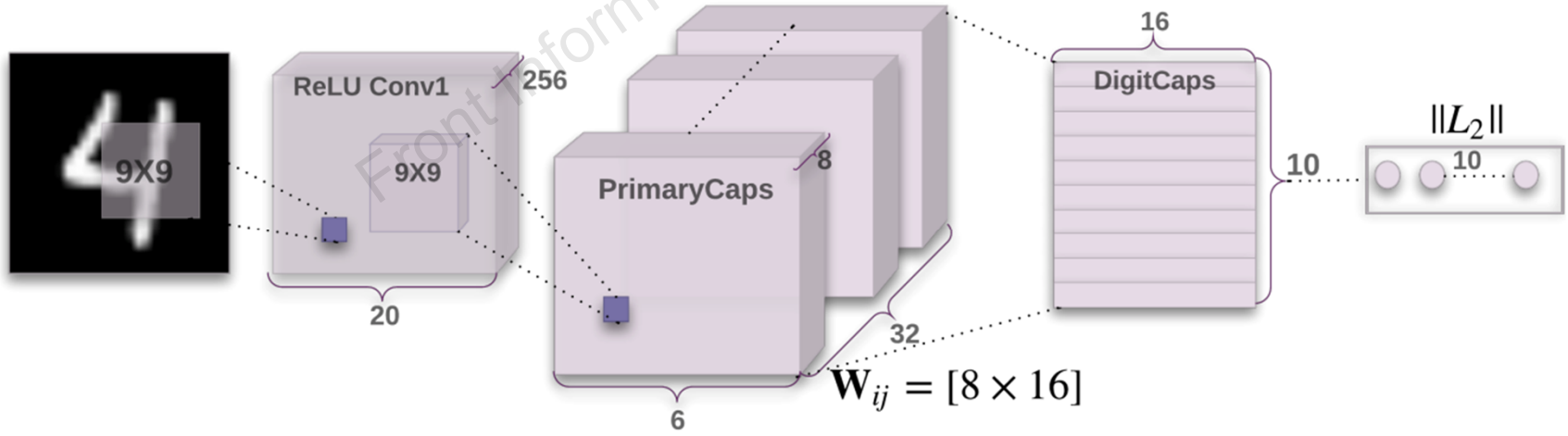
# Visual knowledge in the pre-big-model era: retrospect

## ② Visual relation

*Visual concepts can be related to each other in different ways, resulting in various types of visual relations.*

### 1. Geometric relation

*How objects are arranged and transformed in space, including their position, orientation, size, and shape.*




Capsule network

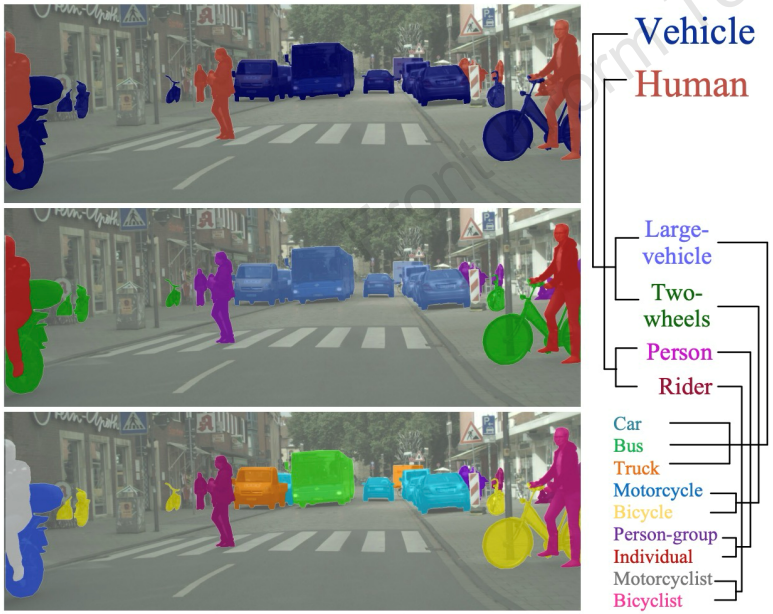
# Visual knowledge in the pre-big-model era: retrospect

## ② Visual relation

*Visual concepts can be related to each other in different ways, resulting in various types of visual relations.*

### 2. Semantic relation

*How objects are related to each other based on their meanings.*




*E.g., a neural parser[2] that can generate structured, pixel-wise descriptions of visual observations in terms of a semantic concept hierarchy.*

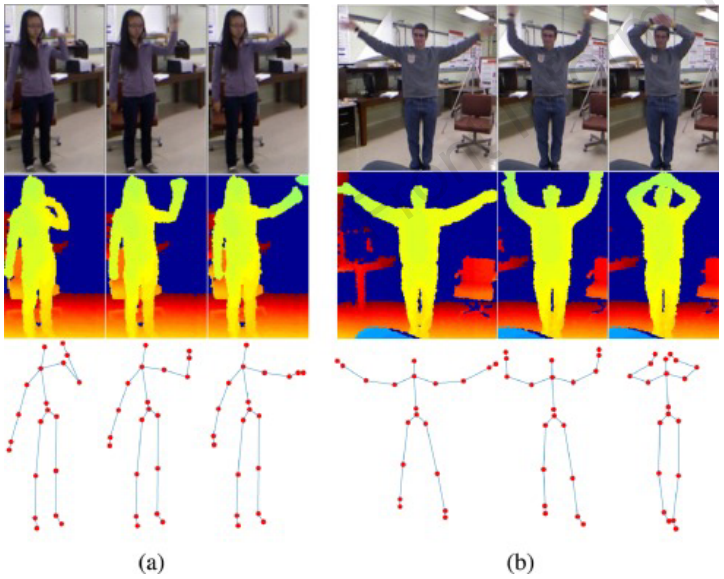
# Visual knowledge in the pre-big-model era: retrospect

## ② Visual relation

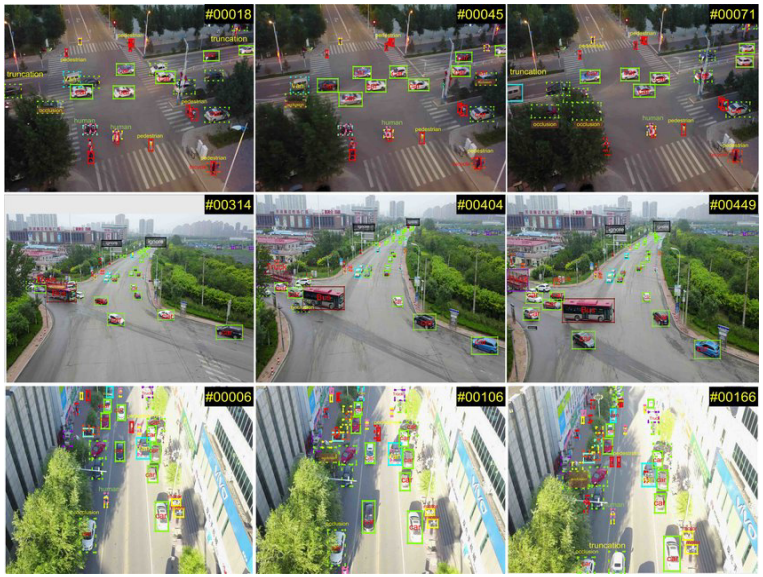
*Visual concepts can be related to each other in different ways, resulting in various types of visual relations.*

### 3. Temporal relation

*Explicate the sequential or chronological order of events and actions as they occur over time within visual data.*




Action recognition




Video object detection

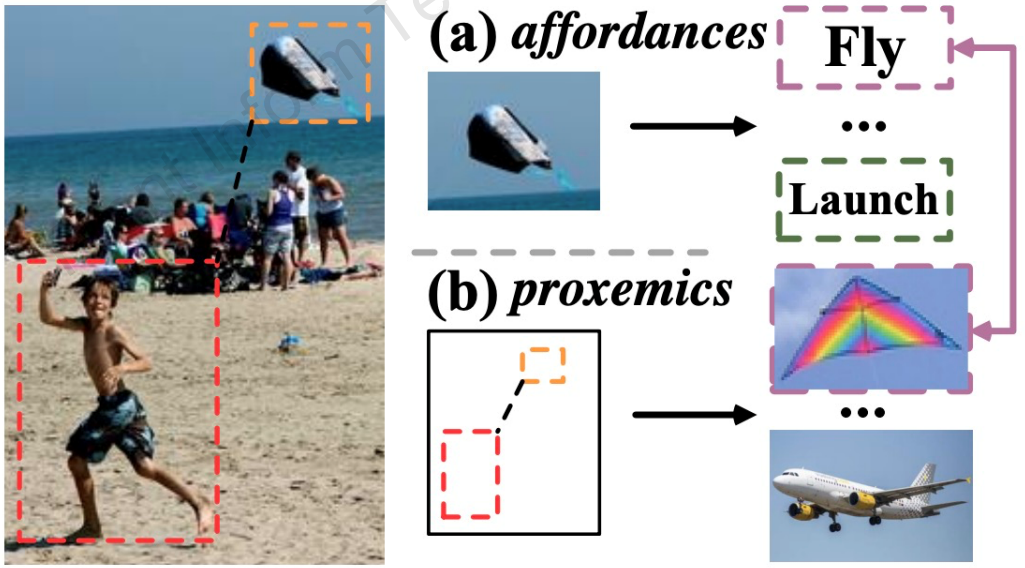
# Visual knowledge in the pre-big-model era: retrospect

## ② Visual relation

*Visual concepts can be related to each other in different ways, resulting in various types of visual relations.*

4. Functional relation

*Actions that objects enable or support.*




Human–object interaction detection[3] & affordance estimation

# Visual knowledge in the pre-big-model era: retrospect

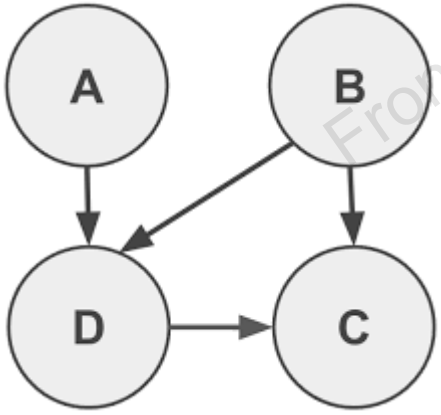
## ② Visual relation

*Visual concepts can be related to each other in different ways, resulting in various types of visual relations.*

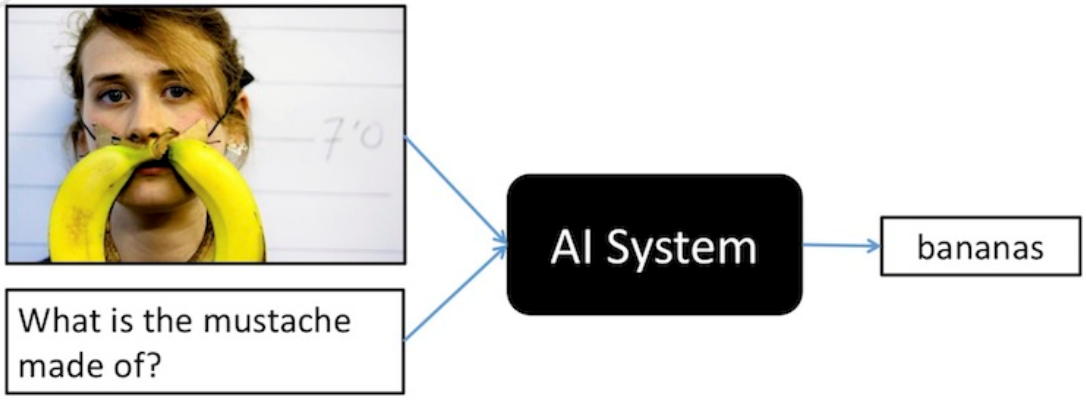
### 5. Causal relations

*How events, actions, or objects within a visual context can directly influence or result in one another.*






Causal reasoning




Visual question answering (VQA)

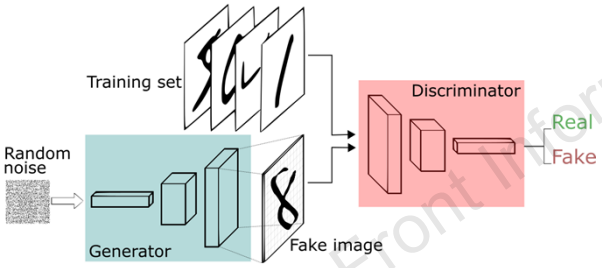
# Visual knowledge in the pre-big-model era: retrospect

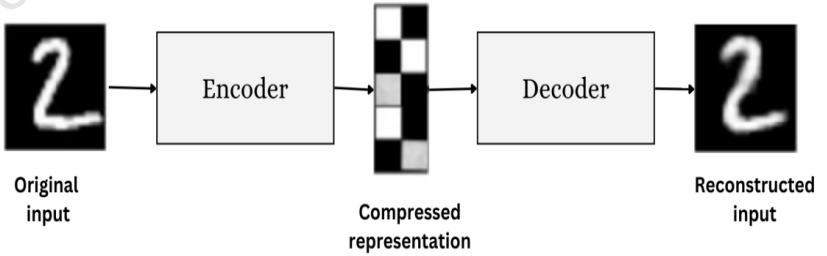
## ③ Visual operation

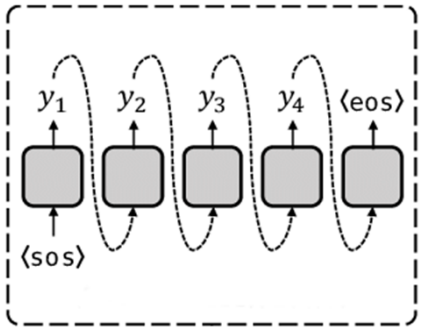
*The manipulation of over visual concepts or objects in space or time.*

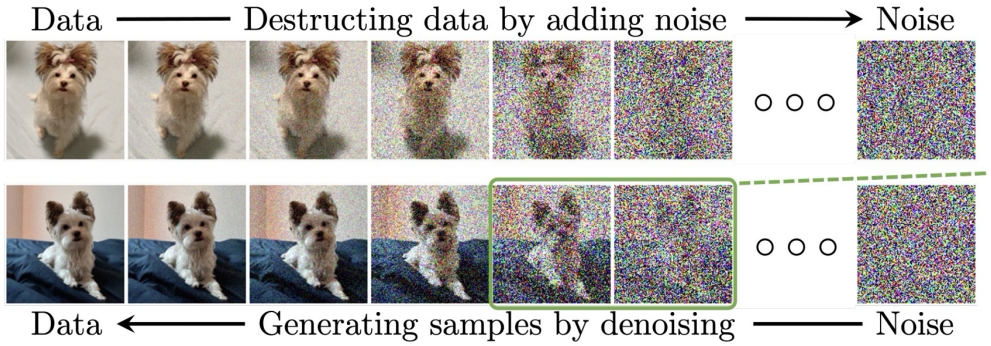
Customized visual content generation

*E.g., GANs, VAEs, autoregressive models, and diffusion models*



# Visual knowledge in the pre-big-model era: retrospect

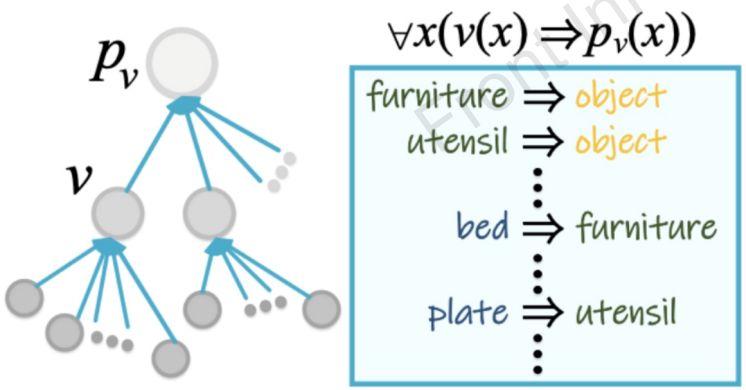
## ④ Visual reasoning

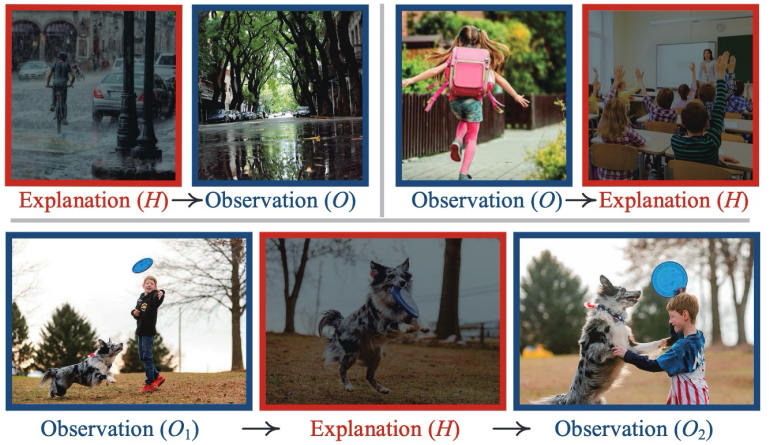
*The process of applying visual concept, visual relation, and visual operation, which are commonalities among various tasks, to draw valid and sound conclusions from premises or evidence.*



Visual question answering (VQA) & visual semantic parsing

*E.g., NeSy-based visual parser[4][5], visual abductive reasoning[6], and LLMs*



# Visual knowledge in the big model era: prospect

## ☐ Empowering big models with visual knowledge

1. **Transparency**: *the degree to which the internal workings and outputs of models can be understood and explained by humans.*

→ The integration of visual knowledge may endow big models with promising ad-hoc interpretability. A notable evidence is the groundbreaking study of deep nearest centroids[7], a fully end-to-end, prototype-based neural classifier.

2. **Reasoning**: *grasping the underlying logic or truth of the content it produces.*

→ Visual knowledge provides an explicit, powerful, and unified framework. Combining big models' implicit knowledge and explicit visual knowledge in a form of multiple knowledge representation[8] is a promising pathway forward.

3. **Catastrophic forgetting**: *the tendency of DNNs to lose their previously learned knowledge when exposed to new data or tasks.*

→ 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, allowing to update knowledge outside the big models, enabling more targeted interventions to prevent catastrophic forgetting.

# Visual knowledge in the big model era: prospect

## ☐ Boosting visual knowledge with big models

1. Big AI models will be an essential cornerstone of visual knowledge

→ It is a natural choice to use the large-scale learning ability of big models to learn robust visual concepts and model basic visual relations.

2. Big AI models will serve as a knowledge source for visual knowledge

→ LLMs have been observed to learn not only contextualized text representations but also a significant body of world knowledge and commonsense knowledge, which suggests the great potential of big models as a knowledge base that could significantly enrich visual knowledge.

3. Big AI models will provide complementary knowledge for visual knowledge

→ The knowledge acquired from the text data not only enriches, but also complements, visual knowledge; i.e., complementing visual knowledge with the knowledge in big models will lead to a more holistic understanding of the world.

# Conclusions

- With data from the Internet and increasingly powerful computing resources, big AI models is able to:

  *- assemble the characteristics of both **connectionism** and the **scaling law**.*

  *- swiftly embed themselves into the fabric of human society, becoming indispensable for scientific discovery.*

- Big AI models still exhibit deficiencies in:

  ***transparency**, **accountability**, and **symbolic reasoning**.*

- Given the 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.

# References

[1] Pan YH, 2019. On visual knowledge. Front Inform Technol Electron Eng, 20(8):1021-1025. https://doi.org/10.1631/FITEE.1910001

[2] Li LL, Zhou TF, Wang WG, et al., 2022. Deep hierarchical semantic segmentation. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1236-1247. https://doi.org/10.1109/CVPR52688.2022.00131

[3] Li LL, Wei JN, Wang WG, et al., 2023b. Neural-logic human-object interaction detection. Proc Int Conf on Neural Information Processing Systems.

[4] Wang WG, Yang Y, Wu F, 2025. Towards data- and knowledge-driven artificial intelligence: a survey on neuro-symbolic computing. IEEE Trans Patt Anal Mach Intell, 47(2):878-899. https://doi.org/10.1109/TPAMI.2024.3483273

[5] Li LL, Wang WG, Yang Y, 2023a. LogicSeg: parsing visual semantics with neural logic learning and reasoning. Proc IEEE/CVF Int Conf on Computer Vision, p.4099-4110. https://doi.org/10.1109/ICCV51070.2023.00381

[6] Liang C, Wang WG, Zhou TF, et al., 2022. Visual abductive reasoning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.15544-15554. https://doi.org/10.1109/CVPR52688.2022.01512

[7] Wang WG, Han C, Zhou TF, et al., 2023. Visual recognition with deep nearest centroids. Proc 11th Int Conf on Learning Representations.

[8] Yang Y, Zhuang YT, Pan YH, 2021. Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies. Front Inform Technol Electron Eng, 22(12):1551-1558. https://doi.org/10.1631/FITEE.2100463