Vol. 16 (2026): Proceedings of the 2026 International Conference on AI and Technological Innovation Applications (AITIA 2026)
E-ISSN: 3105-0522 P-ISSN: 3105-0514
DOI: https://doi.org/10.70267/aitia.2026384389

From Perception to Cognition: A Review of Neural-Symbolic AI for Computer Vision

Luchang He*

Jilin University, Changchun, China

**Corresponding author: Luchang He.*

Abstract

Deep learning has improved core computer vision tasks a great deal, but most perception-based methods still have weak symbol grounding, limited compositional reasoning, poor generalization, and low interpretability. Neural-Symbolic AI combines neural feature learning with symbolic reasoning and has become an active approach for dealing with these problems. This paper reviews recent work on vision-focused Neural-Symbolic AI through the perception-reasoning-cognition pathway. It covers representative architectures, including scene-graph reasoning models, neural modules with differentiable logic, knowledge-graph-enhanced frameworks, and 3D neuro-symbolic grounding methods, along with their applications in visual reasoning and scene understanding. The review examines the main achievements, current limitations, and future prospects of the field. Overall, Neural-Symbolic AI shifts computer vision from pure perception toward structured understanding and reasoning, although problems remain in scalable reasoning, knowledge integration, and real-world deployment. It offers a practical direction for building more explainable and reliable vision systems.

Keywords

neural-symbolic AI, computer vision, visual reasoning, scene understanding, explainability

1. Introduction

Since its emergence in the 1960s, computer vision has shifted from geometric modeling and hand-crafted rules to statistical learning and deep learning. Models such as AlexNet, ResNet, R-CNN, and the Vision Transformer have improved performance in image classification, object detection, and semantic segmentation. Even so, most current vision models remain strongest at the perceptual level, that is, identifying what appears in an image. They are less effective at handling relationships among objects, making judgments that rely on outside knowledge, and explaining how a decision was made.

This gap matters because visual systems used in real-world settings need to do more than detect objects and attributes. They also have to answer higher-level questions such as “why” and “what would happen if the conditions changed?” Existing methods still have clear problems with cross-scene generalization, symbol grounding, and interpretable decision-making [1].

Against this background, neuro-symbolic artificial intelligence has become a major direction in computer vision research. Neural networks are strong at learning representations from high-dimensional images, while

symbolic methods handle knowledge representation, logical reasoning, and interpretability more directly. Combining them could help visual systems move beyond perception toward reasoning and cognition.

Recent studies have shown progress in visual reasoning, scene understanding, knowledge-enhanced vision, and 3D semantic understanding. However, the features, application scenarios, and limitations of different integration paradigms have not been reviewed systematically. This paper therefore reviews advances in neuro-symbolic AI for computer vision through the progression from perception to reasoning and cognition. It examines the main technical paradigms, representative applications, current challenges, and future directions.

2. Overall Development of Neuro-Symbolic AI

Neuro-symbolic artificial intelligence refers to methods that combine the representation-learning ability of neural networks with the knowledge representation, rule modeling, and logical reasoning ability of symbolic AI. Neural networks are good at learning feature representations from high-dimensional data automatically, so they usually perform well on perceptual tasks. Even so, they are still weak in interpretability, use of knowledge, compositional generalization, and stable reasoning. Symbolic methods, by contrast, are stronger in structured representation, logical constraints, and transparent reasoning, but they cannot effectively process complex, continuous raw visual data directly [2].

The main problem in neuro-symbolic AI, then, is not just attaching neural networks to rule-based systems. The harder task is to coordinate perception and reasoning within one architecture. Existing approaches can generally be grouped into three categories. The first uses a symbolic system as the reasoning backbone and a neural network as the perceptual front end. In this setup, images are converted into symbolic representations of objects, attributes, and relations, then passed to a rule-based or knowledge-based system for reasoning. This category is known as neural-to-symbolic.

The second category, symbolic-to-neural, takes a neural network as the main architecture and brings logical rules, relational constraints, or prior knowledge into training. The third puts more emphasis on learning concepts together with compositional reasoning, usually through scene graphs, knowledge graphs, or other intermediate semantic representations. This can be understood as a hybrid intermediate representation. The categories are different in how deeply and in what way they combine the two traditions. Still, they all aim to improve how a system moves from perception to cognition.

This matters especially in computer vision. A visual system must do more than identify what appears in an image. It also needs to capture relations among objects, make inferences with outside knowledge, and support reasoning that can be interpreted. In tasks such as visual question answering, scene understanding, relation recognition, and 3D semantic understanding, models based only on statistical correlations often do not handle complex reasoning well. Neuro-symbolic methods, by comparison, appear to offer a useful direction for higher-level visual understanding and cognition.

3. From Perception to Symbolic Representation

Turning perception into symbols is a necessary middle step because reasoning systems do not work directly on unstructured feature maps. Several approaches have been used for this conversion. Detector-based pipelines rely on object detection or instance segmentation to produce discrete entities with category labels, bounding boxes, masks, and attributes. By contrast, object-centric methods such as Slot Attention divide visual features into slots, so each slot can represent a possible object without depending on a fixed vocabulary. Scene-graph generation builds on these object representations by predicting relational edges such as “left of,” “holding,” or “behind.” Vision-language grounding, meanwhile, links noun phrases and relational expressions to image regions.

These representations vary in the kind of supervision they need and in their level of detail, but they all aim to produce intermediate concepts that later reasoning modules can use again. The point is simple. If one object is missed, if an attribute is assigned incorrectly, or if a relation is predicted wrongly, the output of an otherwise correct symbolic program can change. For that reason, effective neuro-symbolic vision needs accurate perception as well as representations that keep uncertainty visible instead of treating every detected symbol as equally reliable [3-5].

3.1 Neuro-Symbolic Reasoning in Visual Tasks

Once symbolic representations are available, the strengths of neuro-symbolic methods become particularly clear in visual reasoning. A common setup is to turn an image into a structured scene representation, parse a natural-language question into an executable symbolic program, and then run that program over the scene structure to produce an answer [3]. These programs usually use simple operations such as "select," "filter," "compare," and "count." Because of that, the reasoning process is easier to inspect than in end-to-end black-box models.

NS-CL, proposed by Mao et al., gives a clear example of the data efficiency of this approach. Trained on only 5K CLEVR images, less than 10% of the full training set, the method achieved an average concept-recognition accuracy of 98.7% on the diagnostic question set. Even when trained on only 10% of the data, it attained a VQA accuracy of 98.9%, substantially outperforming baseline models such as MAC and TbD [4]. The pattern is fairly clear. By learning object concepts explicitly, along with the compositional rules that connect them, NS-CL relies less on very large labeled datasets. It can then recombine those learned concepts across different problems, which appears to support both sample efficiency and compositional generalization. Even so, these results came mainly from the structurally regular synthetic CLEVR dataset, so it is still uncertain how well the method transfers to more complex real-world scenes.

Program execution is not the only approach. Scene-graph reasoning is another major direction. A scene graph models object-relation-object structures directly, so the model can state more clearly which entities are connected and in what way. The XNM method of Shi et al. shows this clearly: on CLEVR, XNM-GT achieves 100% reasoning accuracy when ground-truth scene graphs and programs are used. Even with detected scene graphs, XNM-Det reaches an overall accuracy of approximately 97.7% [5].

This matters for two reasons. First, scene graphs make the reasoning process easier to interpret. Second, they appear to help with parameter efficiency and generalization to real-world scenes. Program-based and graph-based methods therefore fit together rather than competing with each other. A program defines the sequence of operations required by a question, while a scene graph supplies the entities and relations those operations use. Putting them together separates task interpretation from visual evidence and leaves intermediate states open to inspection. That also makes diagnosis more precise: a wrong answer can be traced to an incorrect object node, relation edge, or program step, instead of being linked only to the final prediction. At the same time, the gap between results with ground-truth and detected scene graphs shows a practical limit. Progress in symbolic reasoning still depends heavily on accurate visual extraction.

3.2 Extending from 2D Vision to 3D Grounding

As work in visual understanding moves from 2D images to 3D environments, neuro-symbolic methods have been used for grounding objects and relations in 3D scenes as well. Language in 3D settings often depends on orientation, viewpoint, and constraints involving several objects at once, so these tasks rely more on structured representations and compositional reasoning. Experimental results for NS3D show that, on the ReferIt3D/SR3D task, it achieves overall accuracies of 62.7% and 62.0% for viewpoint-related questions, outperforming methods such as MVT and BUTD-DETR [6]. The pattern is fairly clear. In data-efficiency experiments, NS3D still reaches 52.0% accuracy when trained on only 1.5% of the data, which suggests that modular, compositional neuro-symbolic structures work well in 3D vision tasks where annotation is expensive.

This kind of 3D grounding is closely connected to embodied AI. A robot has to do more than recognize objects. It also needs to interpret spatial relations such as "the chair to the left of the table" or "the box closest to the door" and turn language instructions into navigation, localization, and manipulation actions. In this setting, object, attribute, and spatial-relation representations from neuro-symbolic methods can link 3D perception, language understanding, and action planning.

Even so, these methods still face several problems. Annotating 3D data remains expensive, and performance can be weakened by sparse point clouds, sensor noise, object occlusion, and viewpoint variation. Errors from front-end perception may then carry over into later symbolic reasoning stages. For practical use in real-world 3D scenes, better perceptual robustness and stronger generalization are still needed.

4. Discussion

4.1 Main Contributions of Neuro-Symbolic AI to Computer Vision

Neuro-symbolic artificial intelligence contributes to computer vision in three main ways. First, it shifts visual systems away from treating images mainly as pixels or low-level features and toward representing objects, attributes, and relations in a more structured form. This lets a model tell apart objects in an image together with their colors, shapes, locations, and relationships to other objects. A model, for instance, can learn the concepts “red,” “cube,” and “to the left of” as separate units and then combine them again to interpret new scenes that are rare or missing in the training data. This makes compositional generalization more feasible.

Second, neuro-symbolic methods support multi-step visual reasoning through program execution, scene graphs, and logical constraints. A model can filter objects that meet a condition, compare their attributes, and count them in sequence. Since learned concepts and rules can be reused across different tasks, these systems often need less training data, and they can learn new concept combinations from fewer examples.

Third, these methods can improve interpretability and controllability. Concept-level representations, relational graphs, and rule-execution paths can show the basis for a model's decisions and help identify whether an error comes from front-end perception or later reasoning [7-9]. That distinction matters.

Their contribution is therefore not limited to better recognition accuracy. They also support more structured understanding, better sample efficiency, stronger compositional generalization, and clearer reasoning processes in visual systems. This reusable structure becomes especially useful when annotations are expensive, because a concept learned in one setting can be carried into another instead of being learned again from raw pixels. In medical imaging, for example, explicit anatomical entities and spatial relations may help with consistency checks. In robotics, reusable object and action concepts can connect perception with planning. Still, these benefits depend on whether the learned symbols remain semantically stable across datasets and environments.

4.2 Current Limitations and Open Problems

Neuro-symbolic methods have clear limitations despite their potential. One problem is the unstable link between visual perception and symbolic representation. Neural networks can generate noisy objects, attributes, and relations, while symbolic reasoning usually needs accurate, structured inputs. When grounding errors occur, later reasoning is affected as well.

A second issue comes from the kinds of evaluation settings still used in many studies. Many experiments depend on controlled datasets and task settings, such as CLEVR, visual question answering, and referring-expression tasks, and these remain quite far from open-ended real-world environments. Systematic reviews have also noted that current research focuses mainly on learning and inference, knowledge representation, and logic and reasoning, while explainability, trustworthiness, and meta-cognition receive insufficient attention [2].

Scalability is another difficulty. As a scene includes more objects and relations, the number of possible symbolic combinations and reasoning paths increases rapidly. Memory and computational costs then rise as well. External knowledge may reduce ambiguity, but using large knowledge graphs creates further problems, including relevance, conflicts, and outdated facts. Most systems are also built around a predefined ontology. If an image contains an unfamiliar object or a relation outside the symbolic vocabulary, the system may fail. For that reason, a practical system needs open-vocabulary concepts, uncertainty-aware reasoning, and incremental updates that do not destabilize previously acquired knowledge.

There is also the issue of reasoning shortcuts. A model may seem to execute a symbolic program or follow logical constraints, yet still produce answers from dataset biases, statistical correlations, or unstable concepts rather than actual reasoning. The problem is similar to hallucination in large language models. An output can look plausible without matching a genuine reasoning process. Existing research has shown that neuro-symbolic systems may suffer from knowledge forgetting, concept bias, and unreliable reasoning paths during continual learning [10]. So neuro-symbolic methods are not automatically the same as trustworthy AI. Their reliability still needs to be tested through causal analysis, counterfactual testing, and evaluation in open-world settings.

4.3 Future Directions

Future research should move toward visual reasoning that can be checked directly. One clear need is for stronger visual-symbol extraction modules that combine open-vocabulary detection, scene-graph generation, and 3D reconstruction. These tools could convert images into structured representations of object categories, spatial locations, attributes, actions, and relations. Another step is to develop executable reasoning programs, so models can parse questions into operations such as filter, relate, count, and compare, then carry them out step by step over object sets or scene graphs. That matters. It would help show whether an error comes from visual recognition, relation extraction, or logical reasoning.

Visual tasks should also bring in causal and counterfactual reasoning. For example, asking whether the answer changes when an object is removed can test whether a model understands causal relationships in a scene, instead of depending on superficial correlations. These programs should also provide confidence estimates for intermediate results. Rather than forcing a single hard symbol at every stage, a system could keep several plausible objects or relations and update them when later evidence becomes available. This may reduce error propagation and better calibrate the final answer [11].

Knowledge graphs can also provide visual commonsense. They can associate “knife,” “cutting board,” and “vegetables” with a cooking scene, or connect “red light,” “crosswalk,” and “pedestrian” to traffic rules. At the same time, more targeted benchmarks are needed, especially ones that include intermediate reasoning steps, out-of-distribution tests, compositional generalization tests, and counterfactual questions. In high-risk applications such as medical imaging, autonomous driving, and robotic manipulation, the need for interpretable and verifiable visual reasoning is even stronger.

The next priority, then, is not just adding symbolic rules. It is building a visual intelligence pipeline that combines strong perception, structured representation, program execution, causal analysis, and knowledge-based reasoning. Multimodal foundation models may work as flexible interfaces between raw perception and symbolic knowledge, but their outputs should be checked and constrained rather than treated as facts. One useful design would let a foundation model propose candidate concepts or programs, then use an executable symbolic component to verify whether they are consistent with the scene.

Evaluation should report more than final-task accuracy. It should also cover grounding accuracy, program validity, calibration, robustness to perturbations, and the faithfulness of explanations. A multi-level evaluation like this would make it easier to tell whether an improvement comes from actual reasoning or from exploiting benchmark regularities.

Overall, neuro-symbolic AI gives computer vision a more structured direction than pure deep learning. It can improve object-relation modeling, multi-step reasoning, and model interpretability, but the current difficulties remain clear, including unstable grounding, reasoning shortcuts, and limited generalization to open-world scenes. As differentiable logic, causal reasoning, continual learning, and multimodal foundation models continue to develop, neuro-symbolic methods will likely take on a larger role in more complex visual systems.

5. Conclusion

This paper reviewed recent work on neuro-symbolic AI in computer vision. The literature suggests that neuro-symbolic methods can bring together the perceptual strengths of neural networks and the reasoning strengths of symbolic approaches. In computer vision, this allows systems to move beyond object recognition toward relational understanding, logical inference, and scene cognition. Reported results also suggest gains in data efficiency, compositional generalization, and interpretability across tasks such as visual question answering, concept learning, scene-graph reasoning, and 3D grounding.

Even so, several problems remain. Current work still runs into difficulties with symbol grounding, generalization in real-world scenes, reasoning stability, and building unified frameworks across tasks. The main issue is continuity. A system may perform well on a final benchmark while still depending on weak links between visual input, concept formation, and symbolic reasoning.

Future work should connect visual representation, symbolic reasoning, causal analysis, and knowledge modeling more directly. If the aim is to move computer vision from perceptual systems toward more reliable

and interpretable cognitive systems, evaluation cannot focus only on final-answer accuracy. It also needs to test the full chain, from perception to grounded concepts and verifiable decisions.

References

[1] D. Yu, B. Yang, D. Liu, H. Wang, and S. Pan, "A Survey on Neural-Symbolic Learning Systems," arXiv preprint arXiv:2111.08164, 2021.

[2] B. C. Colelough and W. Regli, "Neuro-Symbolic AI in 2024: A Systematic Review," in CEUR Workshop Proceedings, vol. 3819, 2024, pp. 1-19.

[3] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding," in NeurIPS, 2018, pp. 1031-1042.

[4] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu, "The Neuro-Symbolic Concept Learner," in ICLR, 2019.

[5] J. Shi, H. Zhang, and J. Li, "Explainable and Explicit Visual Reasoning Over Scene Graphs," in CVPR, 2019, pp. 8376-8384.

[6] J. Hsu, J. Mao, and J. Wu, "NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations," in CVPR, 2023, pp. 2614-2623.

[7] S. Amizadeh et al., "Neuro-Symbolic Visual Reasoning: Disentangling 'Visual' from 'Reasoning'," in ICML, 2020, pp. 279-290.

[8] S. Badreddine, A. d'Avila Garcez, L. Serafini, and M. Spranger, "Logic Tensor Networks," Artificial Intelligence, vol. 303, p. 103649, 2022.

[9] W. Stammer, P. Schramowski, and K. Kersting, "Right for the Right Concept," in CVPR, 2021, pp. 3619-3629.

[10] E. Marconato et al., "Neuro-Symbolic Continual Learning," in ICML, 2023, pp. 23915-23936.

[11] P. Barbiero et al., "Interpretable Neural-Symbolic Concept Reasoning," in ICML, 2023, pp. 1801-1825.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access