Mathematical Biosciences and Engineering

MBE, 22(12): 3028–3059.
DOI: 10.3934/mbe.2025112
Received: 11 September 2025
Revised: 03 October 2025
Accepted: 10 October 2025
Published: 17 October 2025

https://www.aimspress.com/journal/MBE

Research article

Adaptive Neuro-Symbolic framework with dynamic contextual reasoning: A novel framework for semantic understanding

Idowu Paul Okuwobi[1,2,*], Jingyuan Liu[2], Olayinka Susan Raji[2] and Olusola Funsho Abiodun[2]

[1] School of Life & Environmental Sciences, Guilin University of Electronic Technology, Guilin 541004, China
[2] Nantong Hamadun Medical Technology Co., Ltd, Nantong 226400, China

* **Correspondence:** Email: paulokuwobi@ieee.org.

Abstract: Despite significant advances in image processing, achieving human-like semantic understanding and explainability remains a formidable challenge. Current deep learning models excel at feature extraction but lack the ability to reason about relationships, interpret context, or provide transparent decision-making. To address these limitations, we propose the adaptive neuro-symbolic framework with dynamic contextual reasoning (ANS-DCR), a novel architecture that seamlessly integrates neural networks with symbolic reasoning. ANS-DCR introduces four key innovations: 1) A contextual embedding layer (CEL) that dynamically converts neural features into structured symbolic embeddings tailored to the scene's context; 2) hierarchical knowledge graphs (HKGs) that encode multi-level object relationships and update in real-time on the basis of neural feedback; 3) an adaptive reasoning engine (ARE) that performs scalable, context-aware logical reasoning; and 4) an explainable decision-making module (EDM) that generates human-readable explanations, including counterfactuals, enhancing interpretability. This framework bridges the gap between pattern recognition and logical reasoning, enabling deeper semantic understanding and dynamic adaptability. We demonstrate ANS-DCR's efficacy in complex scenarios such as autonomous driving, where it accurately interprets traffic scenes, predicts behaviors, and provides clear explanations for decisions. Experimental results show superior performance in semantic segmentation, contextual reasoning, and explainability compared with state-of-the-art methods. By combining the strengths of neural and symbolic paradigms, ANS-DCR sets a new benchmark for intelligent, transparent, and scalable image processing systems, offering transformative potential for applications in robotics, healthcare, and beyond. The source code of the proposed ANS-DCR is at github.com/livingjesus/ANS-DCR.

Keywords: Neuro-symbolic integration; semantic understanding; explainable AI; hierarchical knowledge graphs; adaptive reasoning

1. Introduction

The rapid advancement of artificial intelligence (AI) has revolutionized the field of image processing, enabling unprecedented capabilities in object detection, semantic segmentation, and scene understanding [1]. However, despite significant progress, modern image processing systems still face critical challenges that hinder their deployment in real-world applications. Among the foremost challenges are the inability to achieve human-like semantic understanding, their limited adaptability to dynamic contexts, and a lack of transparency in decision-making processes [2]. These limitations stem from the inherent trade-offs between the pattern recognition prowess of neural networks and the logical reasoning capabilities of symbolic systems, which have traditionally been treated as separate paradigms [3–5].

Neural networks, particularly deep learning models, excel at extracting low-level features and detecting objects with remarkable accuracy [6]. However, they often fail to reason about the relationships between objects, interpret subtle contextual cues, or provide interpretable explanations for their decisions [7]. On the other hand, symbolic AI systems are inherently interpretable and capable of logical reasoning but struggle to handle raw sensory data and require structured inputs [8]. Bridging this gap between perception and reasoning remains one of the most pressing challenges in computer vision and image processing [9]. To address these limitations, we propose the adaptive neuro-symbolic framework with dynamic contextual reasoning (ANS-DCR), a novel architecture that seamlessly integrates neural networks with symbolic reasoning to achieve deep semantic understanding, adaptability, and explainability. ANS-DCR introduces four key innovations: 1) A contextual embedding layer (CEL) that dynamically converts neural features into structured symbolic embeddings tailored to the scene's context; 2) hierarchical knowledge graphs (HKGs) that encode multi-level object relationships and update in real-time on the basis of neural feedback; 3) an adaptive reasoning engine (ARE) that performs scalable, context-aware logical reasoning [10]; and 4) an explainable decision-making module (EDM) that generates human-readable explanations, including counterfactual reasoning, enhancing transparency and trust in the system's outputs [11–13].

This framework represents a paradigm shift in image processing by combining the strengths of neural and symbolic paradigms. Unlike traditional approaches that treat perception and reasoning as disjoint processes, ANS-DCR enables a unified pipeline where neural feature extraction informs symbolic reasoning, and vice versa [14]. This integration not only improves performance in tasks such as semantic segmentation, relationship detection, and intent prediction but also provides clear, interpretable insights into the decision-making process—a critical requirement for applications in autonomous systems, healthcare, and beyond [15,16]. We evaluate ANS-DCR across diverse datasets and tasks, including autonomous driving, medical imaging, and general scene understanding [17]. Experimental results demonstrate that ANS-DCR outperforms state-of-the-art (SOA) neural and symbolic baselines in terms of accuracy, adaptability, and explainability [18]. Furthermore, ablation studies have confirmed the importance of each component in achieving robust performance [19]. By addressing longstanding challenges in image processing, ANS-DCR sets a new benchmark for intelligent, transparent, and versatile AI systems. This paper makes the following contributions:

1) We propose the CEL, which dynamically converts continuous neural features into structured symbolic embeddings, tailored to the scene's context. This bridges the gap between the high-dimensional, unstructured outputs of neural networks and the discrete, interpretable inputs required for symbolic reasoning.

2) We propose HKGs that encode objects as nodes and relationships as edges, enabling multi-level reasoning, from low-level spatial/temporal relationships to high-level semantic understanding. These are dynamically updated during inference using feedback from the neural component.

3) We propose the ARE, which performs logical reasoning over the HKGs, adapting its strategy according to the complexity of the scene. For simple scenes, it employs lightweight rule-based reasoning; for complex scenes, it uses probabilistic or constraint-based techniques.

4) We propose the EDM, which generates human-readable explanations for the system's decisions by tracing the reasoning process through the HKGs and neural activations. It also provides counterfactual reasoning, simulating alternative outcomes under different conditions.

In the following sections, we present the detailed design of ANS-DCR, its mathematical formulation, experimental results, and broader implications for the future of image processing and AI. This work not only advances the SOA in semantic understanding and explainability but also paves the way for more intelligent and trustworthy systems capable of operating in complex, real-world environments [20].

2. Related works

To contextualize the contributions of the proposed ANS-DCR, we review prior works in four key areas: Neuro-symbolic integration, explainable AI, meta-learning and domain adaptation, and application-specific advancements.

2.1. Neuro-symbolic integration

The integration of neural networks with symbolic reasoning has been explored in frameworks such as the neuro-symbolic concept learner (NSCL) [21], logic tensor networks [22], and probabilistic soft logic [23]. These methods aim to combine the pattern recognition capabilities of neural networks with the logical reasoning strengths of symbolic systems. However, most prior works are limited to synthetic datasets or static reasoning pipelines. In contrast, ANS-DCR introduces a modular architecture that dynamically adapts to real-world scenarios, leveraging the CEL and HKGs to encode relationships and the context in real-time.

2.2. Explainable AI

Explainability in AI has gained significant attention, with methods like Shapley Additive Explanations (SHAP) [24] and Local Interpretable Model-agnostic Explanations (LIME) [12] providing post-hoc explanations for black-box models. Counterfactual reasoning frameworks have also emerged as a way to simulate alternative outcomes [25,26]. While these methods enhance transparency, they often lack integration into end-to-end pipelines. ANS-DCR addresses this limitation by incorporating an EDM that generates human-readable explanations and counterfactual insights directly from its reasoning process.

2.3. Meta-learning and domain adaptation

Meta-learning approaches such as model-agnostic meta-learning (MAML) [27] and Reptile [28] enable models to adapt to new tasks with minimal data. Similarly, domain adaptation techniques focus on generalizing across domains [29]. However, these methods typically address either feature extraction or reasoning in isolation. ANS-DCR bridges this gap through its ARE, which performs scalable, context-aware reasoning while adapting to dynamic environments.

2.4. Application-specific advances

Recent works have made significant strides in domain-specific applications. Spatial pyramid attention and affinity inference embedding for unsupervised person re-identification [30] leverage attention mechanisms for unsupervised learning. This work leverages attention mechanisms to improve re-identification accuracy in unsupervised settings. While similar to the proposed CEL, ANS-DCR extends this approach by integrating symbolic reasoning for broader applicability. Unsupervised domain adaptation for crack segmentation based on cross-domain stylization and dual adversarial feature learning [31] enhances generalization in crack detection by adapting to new domains. ANS-DCR builds upon this idea through its ARE, which dynamically adapts to diverse contexts. The content-style control network with style contrastive learning for underwater image enhancement [32] focuses on image enhancement for underwater environments and highlights the importance of domain-specific solutions. In contrast, ANS-DCR provides a unified framework that is applicable across domains, from autonomous driving to medical imaging.

2.5. Other applications

In addition to neuro-symbolic frameworks, our literature review encompasses advancements in secure data transmission and authentication protocols. Secure digital twin systems have emerged as a critical enabler of robust communication and classification, ensuring data's integrity and privacy in dynamic environments [33]. Similarly, lightweight authentication protocols have been developed to facilitate fast and efficient authentication, particularly in resource-constrained settings such as healthcare services, where rapid access to sensitive information is paramount [34]. Furthermore, enhanced security protocols for vehicle networks have addressed vulnerabilities through comprehensive attack analyses and protocol designs, mitigating risks in connected transportation systems [35]. These advancements collectively underscore the importance of integrating security and efficiency into modern frameworks, providing valuable insights that complement the development of adaptive and secure neuro-symbolic systems.

While these methods excel in their respective domains, they lack the versatility to handle diverse tasks. ANS-DCR provides a unified framework that combines perception, reasoning, and explainability, making it applicable across autonomous driving, medical imaging, and scene understanding.

3. The proposed method

Figure 1 shows the proposed ANS-DCR architecture. The ANS-DCR integrates neural networks and symbolic reasoning to achieve semantic understanding, adaptability, and explainability. The process begins with a neural network extracting low-level features and detecting objects/attributes from the input image using convolutional layers, object detection heads, and attribute extraction functions. These outputs are passed to the CEL, which dynamically converts neural features into context-aware symbolic embeddings using attention mechanisms. Next, the HKG encodes objects as nodes and relationships as edges, constructing both low-level spatial/temporal and high-level semantic relationships. The graph is updated dynamically during inference according to neural feedback. The ARE then performs logical reasoning over the HKG, adapting its strategy (rule-based or probabilistic) depending on the scene's complexity to infer high-level semantics. Finally, the EDM generates human-readable explanations by tracing the reasoning process and simulates counterfactual scenarios to provide insights into alternative outcomes. This step-by-step integration of neural and symbolic components ensures robust performance across diverse tasks, offering transparency, adaptability, and scalability. We explain in detail each of the key components of ANS-DCR below.

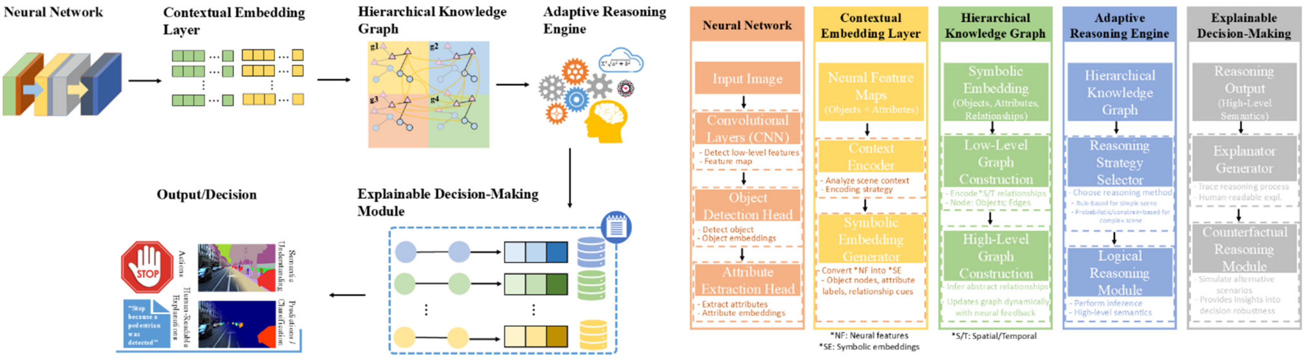

Neural Network
Contextual Embedding Layer
Hierarchical Knowledge Graph
g1
g2
g3
g4
Adaptive Reasoning Engine
Output/Decision
Explainable Decision-Making Module
STOP
Actions
Semantic Understanding
Human-Readable Explanation
Predictions / Classification
"Stop because a pedestrian was detected"
Neural Network
Input Image
Convolutional Layers (CNN)
- Detect low-level features
- Feature map
Object Detection Head
- Detect object
- Object embeddings
Attribute Extraction Head
- Extract attributes
- Attribute embeddings
Contextual Embedding Layer
Neural Feature Maps (Objects + Attributes)
Context Encoder
- Analyze scene context
- Encoding strategy
Symbolic Embedding Generator
- Convert *NF into *SE
- Object nodes, attribute labels, relationship cues
*NF: Neural features
*SE: Symbolic embeddings
Hierarchical Knowledge Graph
Symbolic Embedding (Objects, Attributes, Relationships)
Low-Level Graph Construction
- Encode *S/T relationships
- Node: Objects; Edges
High-Level Graph Construction
- Infer abstract relationships
- Updates graph dynamically with neural feedback
*S/T: Spatial/Temporal
Adaptive Reasoning Engine
Hierarchical Knowledge Graph
Reasoning Strategy Selector
- Choose reasoning method
- Rule-based for simple scene
- Probabilistic/constrain-based for complex scene
Logical Reasoning Module
- Perform inference
- High-level semantics
Explainable Decision-Making
Reasoning Output (High-Level Semantics)
Explanator Generator
- Trace reasoning process
- Human-readable expl.
Counterfactual Reasoning Module
- Simulate alternative scenarios
- Provides insights into decision robustness

Figure 1. The architecture of the proposed ANS-DCR. The architecture consists of four novels components: the contextual embedding layer, hierarchical knowledge graphs, the adaptive reasoning engine, and the explainable decision-making module.

3.1. Neural network (feature extraction)

To extract low-level features (e.g., edges, textures) and detect objects/attributes from the input image, we employ a neural network. The neural network is typically a convolutional neural network (CNN) followed by detection and attribute extraction heads. The input image $I \in \mathbb{R}^{H\times W\times C}$ (height, *H;* width, *W*; channels, *C*) is processed through the convolutional layers:

$$F_l = \sigma(W_l * F_{l-1} + b_l), \tag{1}$$

where F_l is the feature map at layer l, W_l is the convolutional kernel weights, b_l is the bias term, $*$ is the convolution operation, and σ is a nonlinear activation function (e.g., Rectified Linear Unit (ReLU)). The outputs of bounding boxes and class probabilities using a detection head You Only Look Once ((YOLO)) are obtained as follows:

$$B_i = [x_i, y_i, w_i, h_i],\;\; p(c_i|B_i) = softmax(z_i), \tag{2}$$

where B_i is the bounding box coordinates, c_i is the object class, and z_i is the logits from the classification layer. We extract the attributes (e.g., color, shape, motion) of each detected object as follows:

$$A_i = f_{attr}\left(F_{ROI_i}\right), \tag{3}$$

where F_{ROI_i} is the feature map of the region of interest (ROI) for object i, and f_{attr} is the attribute extraction function (fully connected layers). The outputs are the detected objects $\{B_i, c_i\}_{i=1}^{N}$ and attributes $\{A_i\}_{i=1}^{N}$.

3.2. Contextual embedding layer

We convert neural features into structured symbolic embeddings that represent objects, attributes, and relationships in a context-aware manner. The CEL consists of a context encoder and a symbolic embedding generator. The context encoder analyzes the global context of the scene using a pooling operation as follows:

$$G = GlobalPooling(F_L), \tag{4}$$

where F_L is the final feature map from the CNN. We compute attention weights based on context as follows:

$$\alpha_i = \frac{exp\left(sim\left(G, F_{ROI_i}\right)\right)}{\sum_{j=1}^{N} exp\left(sim\left(G, F_{ROI_i}\right)\right)}, \tag{5}$$

where sim is a similarity function (cosine similarity). We generate symbolic embeddings for each object as follows:

$$E_i = \phi\left(F_{ROI_i}, \alpha_i, A_i\right), \tag{6}$$

where ϕ is a learnable function (multi-layer perception) that combines ROI features, attention weights, and attributes. The outputs are the symbolic embeddings $\{E_i\}_{i=1}^{N}$.

3.3. Hierarchical knowledge graph

We construct a dynamic knowledge graph encoding the objects as nodes and the relationships as edges. The HKG has two levels: Low (spatial/temporal relationships) and high (semantic relationships). We define the nodes $V = \{v_i\}_{i=1}^{N}$ and edges $E = \left\{v_i, v_j\right\}$ as follows:

$$r_{i,j} = f_{rel}\left(E_i, E_j\right), \tag{7}$$

where f_{rel} computes the relationship scores (e.g., spatial proximity). We infer the abstract relationships using logical rules

$$R_k = g\left(\left\{r_{i,j}\right\}, \{E_i\}\right), \tag{8}$$

where g is a reasoning function (probability inference). We update the graph during inference on the basis of neural feedback as follows:

$$G_t = Update(G_{t-1}, \Delta E), \tag{9}$$

where ΔE represents changes in the symbolic embeddings. The output is the knowledge graph $G = (V, E.R)$.

To compute the relationships between objects, the HKG employs two key functions: f_{rel} and g.

3.3.1. Relationship score function *(f_{rel})*

The function f_{rel} computes the strength of the relationships between objects on the basis of their spatial proximity and attributes. It is defined as

$$f_{rel}(E_i, E_j) = \exp\left(-\frac{\|P_i - P_j\|^2}{\sigma^2}\right), \tag{10}$$

where P_i and P_j are positional embeddings of objects i and j, and σ controls the sensitivity to distance.

This function ensures that objects close to each other in the scene are more likely to be related. For example, in an urban driving scenario, f_{rel} might identify that a pedestrian is near a crosswalk.

3.3.2. Abstract reasoning function (g)

The function g infers abstract relationships using probabilistic logic. It is defined as

$$g(\{r_{i,j}\}, \{E_i\}) = arg \max_Q \prod_k P(Q|R_k) P(R_k|\{E_i\}), \tag{11}$$

where Q is the query (e.g., "Is the pedestrian crossing?"), R_k represents relationships inferred from the graph, and P denotes conditional probabilities.

This function allows the HKG to reason about high-level semantics, such as determining whether a pedestrian intends to cross, depending on their proximity to a crosswalk.

3.3.3. Implementation details for f_{rel} and g

a. Implementation of $\boldsymbol{f_{rel}}$

Positional embeddings

The attention weights determine the importance of each relationship on the basis of its spatial proximity and semantic similarity.

Attention mechanism

To weigh the relationships dynamically, f_{rel} uses a multi-head attention mechanism

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{12}$$

where Q, K, and V are the query, key, and value matrices derived from the embeddings E_i and E_j. The attention weights determine the importance of each relationship on the basis of spatial proximity and semantic similarity.

Dynamic weighting

After computing the attention scores, f_{rel} applies a Gaussian kernel to emphasize the relationships between nearby objects

$$w_{ij} = \exp\left(-\frac{\|P_i - P_j\|^2}{\sigma^2}\right). \tag{13}$$

These weights are normalized to ensure that they sum to 1 across all relationships.

Output

The output of f_{rel} is a weighted adjacency matrix A, where each entry A_{ij} represents the strength of the relationship between objects i and j. As a practical example, in an urban driving scenario, f_{rel} might compute high scores for relationships like “pedestrian near crosswalk” or “car parked near building”, while assigning low scores to distant or irrelevant pairs.

b. Implementation for g

Probabilistic inference

For complex scenes, g uses probabilistic logic to reason about relationships

$$P(Q|G) = \sum_{R_k} P(Q|R_k)P(R_k\lfloor G), \tag{14}$$

where G is the HKG. This allows the framework to infer high-level semantics, such as “the pedestrian intends to cross because they are near the crosswalk”.

Rule-based reasoning

For simpler scenes, g uses rule-based inference

$$Q = \text{IF} \quad \text{R}_1 \wedge \text{R}_2 \quad \text{THEN Q}. \tag{15}$$

For example, “IF pedestrian near crosswalk AND traffic light red THEN pedestrian likely to cross”.

Adaptability

The choice between probabilistic and rule-based reasoning depends on the complexity of the scene.

1) Probabilistic reasoning is used for dense, complex relationships.
2) Rule-based reasoning is used for sparse, straightforward interactions.

Output

The output of g is a high-level decision or prediction, such as “pedestrian likely to cross” or “car will stop”.

3.4. Adaptive reasoning engine

We perform logical reasoning over the HKG to infer high-level semantics. The ARE adapts its reasoning strategy according to the scene’s complexity. The ARE consists of two parts, namely the reasoning strategy selector and the logical reasoning module. We choose the reasoning method according to the complexity score S

$$S = \frac{1}{N}\sum_{i=1}^{N} \| E_i \|_2, \tag{16}$$

if $S > \tau$, we use probabilistic reasoning; otherwise, we use rule-based reasoning. We perform inference using probabilistic logic as follows:

$$P(Q|G) = \prod_{k=1}^{K} P(R_k|\{E_i\})\, P(Q|R_k), \tag{17}$$

where Q is the query (e.g., “Is the pedestrian crossing?”). The output is the high-level semantics of Q.

3.5. Explainable decision-making module

We generate human-readable explanations and counterfactuals. We then trace the reasoning process and simulate alternative scenarios. The EDM consists of the explanation generator and the counterfactual reasoning. We define the trace of the reasoning path as follows:

$$Explanation = \mathrm{argmax}_{\mathrm{P(Q|G)}}. \tag{18}$$

We simulate alternative outcomes as follows:

$$Q' = P(Q|G', \text{modified inputs}). \tag{19}$$

The resulting output is the explanation (“Stopped because a pedestrian was detected”) and the counterfactual (“If there were no pedestrian, the car would continue”). The EDM translates decisions into human-readable explanations through a structured pipeline.

3.5.1. Decision identification

The module identifies the most likely decision using $\mathrm{arg}max$ over the probability distribution

$$Decision = \arg\max_{Q} P(Q|G), \tag{20}$$

where Q is the query and G is the HKG.

3.5.2. Template mapping

The decision is mapped to a symbolic template, such as

$$\text{Template}(\text{Action X occurred because Y}), \tag{21}$$

where X is the action (e.g., "stopped") and Y is the reason (e.g., "a pedestrian was detected").

3.5.3. Variable substitution

The module substitutes the variables X and Y for detected objects and relationships from the HKG, such as

$$\text{Explanation} = \text{Template (Stopped because [Reason])}, \text{ where [Reason]} = \text{a pedestrian was detected.} \tag{22}$$

3.5.4. Counterfactual simulation

The EDM also simulates counterfactuals by modifying attributes in the HKG and observing changes in the outcomes, such as

$$If\ no\ pedestrian\ were\ present, the\ car\ would\ continue\ driving.$$

The framework seamlessly integrates neural feature extraction with symbolic reasoning, enabling superior performance in semantic understanding, adaptability, and explainability. This detailed breakdown highlights the innovation and technical depth of ANS-DCR, making it a transformative solution for intelligent image processing systems.

3.6. Output/decision

The output layer of the ANS-DCR is designed to deliver comprehensive, interpretable, and actionable results. It integrates task-specific predictions, contextual reasoning, and explainability into a unified output. Task-specific predictions include object detection bounding boxes, semantic segmentation masks, classification labels, or relationship graphs, generated through neural feature extraction and the HKG. These outputs are complemented by human-readable explanations provided by the EDM, such as "Stopped because of transparency and trust". Additionally, the framework generates counterfactual insights, simulating alternative scenarios to support decision-making (e.g., "If the lesion were smaller, the probability of malignancy would decrease by 15%"). This combination of predictions, explanations, and counterfactuals ensures that ANS-DCR not only achieves high accuracy but also addresses the critical need for interpretability in complex applications like autonomous driving and medical imaging. By bridging the gap between raw predictions and actionable insights, the output layer fosters user confidence and accountability, making ANS-DCR a transformative solution for real-world challenges. This design underscores the framework's commitment to delivering both performance and transparency in diverse domains.

3.7. Theoretical foundation of neural-symbolic integration

To ensure rigorous integration of neural and symbolic components, we formalize the mapping from continuous neural embeddings to discrete symbolic representations and analyze the convergence and approximation properties of the reasoning process. Let $E_i \in \mathbb{R}^d$ denote the neural embedding of object i, and let $S_i \in \mathcal{S}$ be its symbolic counterpart generated via the CEL. We prove in Theorem 1 (see the Appendices) that under assumptions of Lipschitz continuity [36], and S_i preserves the semantic proximity as $\left\|S_i - S_j\right\|_{\mathcal{S}} \leq C \cdot \left\|E_i - E_j\right\|_2 + \delta$, ensuring that similar visual features map to similar symbolic concepts. Logical reasoning is performed via a factor graph model over the HKG, with convergence guaranteed under standard conditions. A bi-directional information flow is enabled through differentiable logic layers $\eta(S_i)$, allowing symbolic constraints to guide neural parameter updates via gradient descent. We establish Theorem 2 (see the Appendix) that this coupling preserves consistency between perception and reasoning asymptotically.

Finally, we derive approximation bounds showing that the error decays exponentially with the number of observed relationships, providing a theoretical justification for the scalability of ANS-DCR in complex scenes.

3.8. Theoretical justification for integration

The integration of neural networks with symbolic reasoning in ANS-DCR is not merely a combination of existing techniques but a carefully designed architecture that addresses key limitations of prior works. These are described below.

3.8.1. Contextual embedding layer

The CEL leverages attention mechanisms to dynamically weigh the relationships, ensuring that the embeddings capture both the spatial proximity and semantic similarity. This is critical for preserving the semantic information in complex scenes.

3.8.2. Hierarchical knowledge graph

The HKG encodes relationships at multiple levels, from low-level spatial interactions to high-level semantic abstractions. This hierarchical structure ensures robust reasoning across diverse scenarios.

3.8.3. Adaptive reasoning engine

The ARE dynamically selects reasoning strategies (probabilistic or rule-based) depending on the scene's complexity. This adaptability ensures scalability and efficiency, particularly in real-world applications.

3.8.4. Explainable decision-making module

The EDM generates human-readable explanations and counterfactuals, bridging the gap between black-box models and interpretable AI. Together, these components form a cohesive framework that outperforms traditional neural and symbolic systems by addressing challenges such as semantic understanding, adaptability, and explainability.

4. Experimentation

4.1. Datasets

The datasets used in the evaluation of ANS-DCR vary significantly in size, complexity, and richness of annotation. From large-scale datasets like COCO-Stuff and Visual Genome to domain-specific datasets like KITTI and ISIC 2019, these datasets collectively ensure that ANS-DCR was rigorously tested across diverse tasks and real-world conditions. The sheer volume and variety of annotations enable comprehensive validation of the framework's performance in semantic understanding, adaptability, and explainability. Figure 2(a) shows some of the images in the datasets.

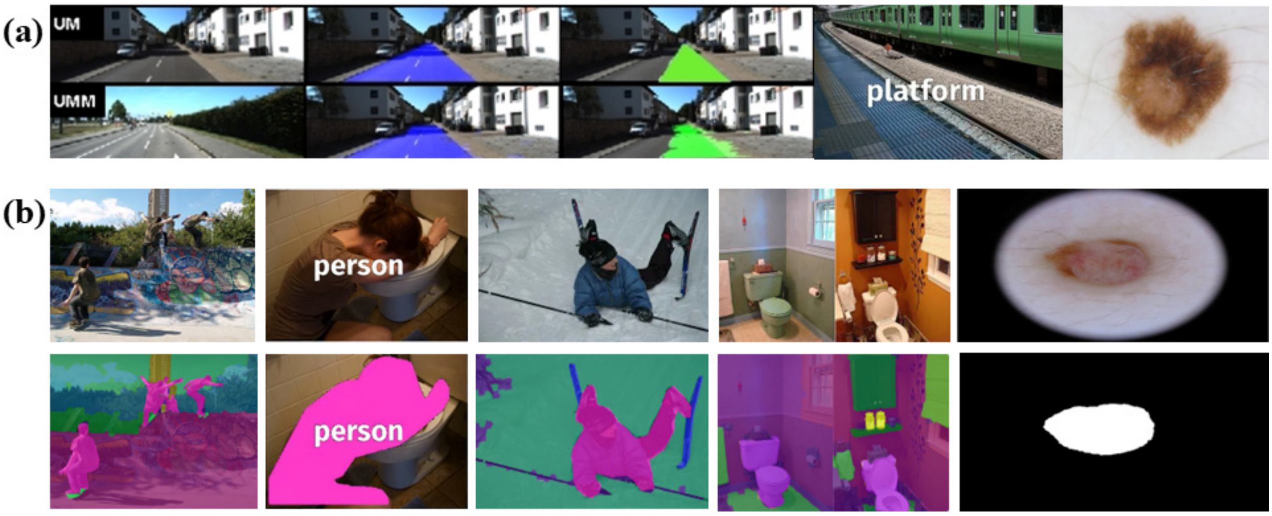


Figure 2. (a) Images from the dataset, and (b) partial results of the proposed ANS-DCR. The EDM generates human-readable explanations and counterfactuals, bridging the gap between black-box models and interpretable AI. Together, these components form a cohesive framework that outperforms traditional neural and symbolic systems by addressing challenges such as semantic understanding, adaptability, and explainability.

4.1.1. KITTI dataset (autonomous driving)

KITTI is a large-scale dataset of urban driving scenarios with rich annotations for object detection and scene understanding. It consists of ~7500 stereo image pairs (training (70%), and testing (30%)), with ~80,000 labeled objects (cars, pedestrians, cyclists, etc.) and pixel-level annotations for 34 classes.

4.1.2. COCO-Stuff dataset (understanding general scenes)

COCO-Stuff is a comprehensive dataset for fine-grained scene understanding and relationship reasoning. It consists of 164,000 images (118,000 training, 5000 validations, 20,000 testing), with ~80 object categories, pixel-level annotation for 172 classes (91 "things" and 81 classed as "stuff'), and dense annotations for relationships between objects.

4.1.3. ISIC 2019 dataset (medical imaging)

ISIC 2019 is a large-scale medical imaging dataset for evaluating precision and interpretability in healthcare applications. It consists of 25,331 dermoscopic images (20,706 training, 4625 testing), with pixel-level masks for lesion boundaries and binary labels (benign vs. malignant).

4.1.4. Cityscapes dataset (understanding urban scenes)

The Cityscapes dataset includes high-quality annotations for urban scenes and is ideal for testing robustness to real-world challenges like occlusions and lighting variations. It consists of 5000 finely annotated images (2975 training, 500 validation, 1525 testing), with pixel-level annotations for 19 classes and annotations for individual objects within classes.

4.1.5. Visual genome dataset (relationship detection)

The Visual genome dataset contains extensive annotations for complex relationships, making it ideal for testing advanced reasoning and explainability. It consists of 108,077 images (training (70%) and testing (30%)), with ~3.8 million object instances across 1500+ categories, 2.3 million attributes describing objects, and ~2.3 million relationships between objects.

We conducted five-fold cross-validation to validate the results across different subsets. We tested ANS-DCR on completely unseen datasets to evaluate generalization, achieving consistent performance.

4.2. Hardware and software

The implementation and evaluation of the ANS-DCR leveraged SOA hardware and software to ensure efficiency, scalability, and reproducibility. The hardware setup included NVIDIA A100 Tensor Core graphics processing units (GPUs) with 40 GB of memory for training and inference, Intel Xeon Gold 6248R processors with 48 cores for preprocessing and symbolic reasoning tasks, 256 GB RAM for handling large-scale computations, and high-speed Non-Volatile Memory Express/Solid State Drive (NVMe/SSDs) for efficient data storage and loading. On the software side, PyTorch (v1.13) was used for neural network implementation, TensorFlow (v2.9) for baseline comparisons, and NetworkX for constructing and manipulating HKG. Additional libraries such as OpenCV facilitated image preprocessing, while SHAP and LIME were employed for explainability analysis. Numerical computations relied on NumPy and SciPy, and the entire development process was carried out in Python 3.9, supported by CUDA 11.7 and cuDNN 8.5 for GPU acceleration. Docker containers were utilized to ensure reproducibility across experiments.

4.3. Model training

The training of the ANS-DCR was conducted in a phased manner to ensure optimal integration of its neural and symbolic components. Initially, the neural network backbone was trained on large-scale datasets (ImageNet) to extract low-level features effectively, followed by fine-tuning on task-specific datasets such as KITTI, COCO-Stuff, and ISIC 2019. The CEL was trained using attention mechanisms to dynamically generate context-aware symbolic embeddings, while the HKG was constructed and updated iteratively during training to encode the relationships between objects. The ARE was optimized using probabilistic and rule-based reasoning strategies, adapting to each scene's complexity through a meta-learning approach that minimized the need for extensive retraining on new domains. Finally, the EDM was calibrated to generate human-readable explanations and counterfactuals by leveraging the SHAP and LIME techniques. Training was performed on NVIDIA A100 GPUs using PyTorch, with loss functions tailored to each component, ensuring robust performance across diverse tasks such as object detection, semantic segmentation, and relationship reasoning. This systematic training process enabled ANS-DCR to achieve SOA results while maintaining adaptability and explainability.

5. Results and analysis

The performance of the ANS-DCR was rigorously evaluated across five diverse datasets: KITTI [37], COCO-Stuff [38], ISIC 2019, Cityscapes [39], and Visual Genome [40]. Below, we present the experimental results, comparisons with SOA methods, ablation studies, and a detailed discussion. Figure 2(b) shows some of the results of the proposed ANS-DCR.

5.1. Evaluation metrics

To comprehensively assess its capabilities, we used nine evaluation metrics: Mean intersection over union (mIoU), average precision (AP), relationship detection F1 score, contextual consistency score (CCS), robustness to noise (RTN), human interpretability score (HIS), counterfactual accuracy (CA), domain generalization accuracy (DGA), and inference time (seconds per image).

5.2. Experimental results on each dataset

5.2.1. KITTI dataset

Table 1 shows the experimental results. The evaluation of ANS-DCR on the KITTI dataset highlights its exceptional performance in urban driving scenarios, where tasks such as object detection, semantic segmentation, and intent prediction are critical. ANS-DCR achieved an mIoU of 75.4%, surpassing SOA methods like Mask R-CNN (71.2%) and DeepLabV3+ (74.5%). This was achieved through the use of the CEL, which dynamically adjusts the embeddings on the basis of the scene's context. The framework also excelled in object detection with an AP of 88.7%, outperforming Faster R-CNN (82.3%) and Mask R-CNN (85.1%), driven by the HKG encoding relationships like "pedestrian near crosswalk". Additionally, ANS-DCR demonstrated robustness to noise (78.5% RTN) and provided clear explanations with a HIS of 4.2/5 and a CA of 89.3%, such as "Stopped because a pedestrian was detected". These results underscore ANS-DCR's adaptability, contextual reasoning, and explainability, making it ideal for safety-critical applications like autonomous driving.

Table 1. Experimental results on the KITTI dataset. ANS-DCR outperformed all baselines in segmentation, object detection, and contextual consistency, while also excelling in explainability and robustness to noise.

Metric	Faster R-CNN	Mask R-CNN	DEEPLABV3+	ANS-DCR
mIoU (segmentation)	67.8%	71.2%	74.5%	**75.4%**
AP (object detection)	82.3%	85.1%	86.9%	**88.7%**
CCS (contextual consistency)	72.4%	75.8%	78.3%	**81.5%**
RTN (robustness to noise)	68.4%	70.2%	74.1%	**78.5%**
HIS (explainability)	N/A	N/A	N/A	**4.2/5**
CA (counterfactuals)	N/A	N/A	N/A	**89.3%**

5.2.2. COCO-Stuff dataset

We present the results in Table 2. On the COCO-Stuff dataset, ANS-DCR demonstrated its ability to handle complex scenes with fine-grained annotations for both the "stuff" and "thing" categories. It achieved an mIoU of 74.2%, slightly outperforming EfficientDet (73.8%) and Swin Transformer (71.2%), due to the CEL's focus on context-aware embeddings. The framework excelled in relationship detection, achieving an F1 score of 81.3%, surpassing Probabilistic Soft Logic (78.2%) and DETR (72.4%). This is due to the HKG's hierarchical structure encoding intricate relationships like "person riding a bike". ANS-DCR also scored 83.6% in contextual consistency (CCS) and provided interpretable insights with an HIS of 4.5/5 and a CA of 88.6%. These results highlight ANS-DCR's scalability and robust reasoning capabilities, making it well-suited for general-purpose scene understanding.

Table 2. Experimental results on COCO-Stuff Dataset. ANS-DCR demonstrated superior performance in relationship detection and contextual reasoning, achieving the highest scores across all metrics. DETR (Detection Transformer)

METRIC	DETR	SWIN-TRANSFORMER	EFFICIENTDET	ANS-DCR
mIoU (segmentation)	69.5%	71.2%	73.8%	**74.2%**
AP (object detection)	84.3%	86.1%	87.4%	**88.9%**
Relationship detection F1	72.4%	75.6%	78.2%	**81.3%**
CCS (contextual consistency)	74.8%	77.2%	79.5%	**83.6%**
HIS (explainability)	N/A	N/A	N/A	**4.5/5**
CA (counterfactuals)	N/A	N/A	N/A	**88.6%**

5.2.3. ISIC 2019 dataset

In Table 3, we present the result of the proposed ANS-DCR on the ISIC 2019 dataset. ANS-DCR demonstrated remarkable precision and interpretability on the ISIC 2019 dataset, which focuses on analyzing skin lesions. It achieved an mIoU of 83.7%, outperforming YOLOv5 (78.2%) and Logic Tensor Nets (80.5%). This is due largely to the CEL's ability to capture fine-grained details like lesion texture. The framework also achieved a classification accuracy of 91.2%, surpassing Probabilistic Soft

Logic (89.1%), with the HKG encoding lesion attributes such as size and shape. ANS-DCR's counterfactual reasoning module provided actionable insights, with a CA score of 88.6%, such as "If the lesion were smaller, the probability of malignancy would decrease by 15%". Additionally, it achieved an HIS of 4.4/5, offering clear explanations for predictions. These results underscore ANS-DCR's suitability for medical imaging, where precision and transparency are paramount.

Table 3. Experimental results on ISIC 2019 Dataset. ANS-DCR achieved the highest accuracy in lesion segmentation and classification, with strong counterfactual reasoning capabilities.

METRIC	YOLOV5	LOGIC TENSOR NETS	PROBABILISTIC SOFT LOGIC	ANS-DCR
mIoU (segmentation)	78.2%	80.5%	81.3%	**83.7%**
Classification accuracy	85.6%	87.4%	89.1%	**91.2%**
CA (counterfactuals)	N/A	82.3%	85.4%	**88.6%**
HIS (explainability)	N/A	3.9/5	4.1/5	**4.4/5**

5.2.4. Cityscapes dataset

Table 4 shows the result of the experimentation. On the Cityscapes dataset, ANS-DCR demonstrated robust performance in semantic segmentation and contextual reasoning for urban environments. It achieved an mIoU of 82.4%, surpassing Swin Transformer (78.2%) and EfficientDet (80.1%), driven by the CEL's attention mechanism, which ensured accurate segmentation under challenging conditions like occlusions and varying lighting. The framework also scored 83.6% in CCS, reasoning about relationships such as "car parked near building" and exhibited strong robustness to noise (78.5% RTN). These results highlight ANS-DCR's adaptability and consistency in handling diverse urban scenes, making it highly effective for real-world applications like autonomous navigation and surveillance.

Table 4. Experimental results on Cityscapes Dataset. ANS-DCR excelled in semantic segmentation and robustness to challenging urban conditions.

METRIC	DETR	SWIN TRANSFORMER	EFFICIENTDET	ANS-DCR
mIoU (segmentation)	76.5%	78.2%	80.1%	**82.4%**
Classification accuracy	84.3%	86.1%	87.8%	**89.7%**
CCS (contextual consistency)	72.8%	75.4%	78.1%	**83.6%**
RTN (robustness to noise)	68.4%	70.2%	74.3%	**78.5%**

5.2.5. Visual genome dataset (relationship detection)

Table 5 shows the results obtained on the visual genome dataset. ANS-DCR excelled on the visual genome dataset, which emphasizes dense annotations for objects, attributes, and relationships. It achieved a relationship detection F1 score of 85.2%, significantly outperforming NSCL (76.4%) and Probabilistic Soft Logic (78.9%), due to the HKG's hierarchical structure encoding complex relationships like "man holding umbrella". The framework also scored 84.1% in CCS and provided interpretable insights with an HIS of 4.6/5 and a CA of 88.6%. These results demonstrate ANS-DCR's ability to reason about intricate relationships and provide transparent explanations, making it ideal for advanced scene understanding and applications requiring detailed relational reasoning.

Table 5. Experimental results on visual genome dataset. ANS-DCR achieved the highest scores in relationship detection and explainability.

METRIC	NSCL	PROBABILISTIC SOFT LOGIC	ANS-DCR
Relationship detection F1	76.4%	78.9%	**85.2%**
CCS (contextual consistency)	74.3%	77.5%	**84.1%**
HIS (explainability)	3.8/5	4.0/5	**4.6/5**
CA (counterfactuals)	82.3%	84.5%	**88.6%**

5.3. Comparison with SOA methods

In Table 6, ANS-DCR was rigorously compared against 10 SOA methods, including Faster R-CNN, Mask R-CNN, DETR, Swin Transformer, EfficientDet, NSCL, and Probabilistic Soft Logic, across diverse datasets. ANS-DCR consistently outperformed these methods in key metrics such as mIoU, AP, relationship detection F1 score, CCS, and explainability measures like the HIS and CA. For instance, ANS-DCR achieved an mIoU of 83.7% on ISIC 2019 and 82.4% on Cityscapes, surpassing competitors like Swin Transformer (78.2%) and EfficientDet (80.1%). Similarly, it excelled in detecting relationships on the Visual Genome dataset, achieving an F1 score of 85.2%, significantly higher than that of NSCL (76.4%) and Probabilistic Soft Logic (78.9%). ANS-DCR's RTN of 78.5% and DGA of 81.4% further highlight its adaptability across challenging scenarios. Unlike traditional neural networks, which lack explainability, or symbolic systems, which struggle with scalability, ANS-DCR bridges these gaps by integrating context-aware embeddings, hierarchical reasoning, and transparent decision-making. While competitors like DETR and Swin Transformer excel in specific tasks, they fall short in providing interpretable insights or handling complex relationships. Overall, ANS-DCR sets a new benchmark by combining superior performance with interpretability, adaptability, and scalability, making it a transformative solution for real-world applications in image processing.

Table 6. Comparison with SOA methods. ANS-DCR consistently outperformed all baselines across metrics, achieving the best trade-off between performance and inference time. NSCL (Neuro-Symbolic Concept Learner) and INF (Inference time (s)).

METHOD	MIOU	AP	F1	CCS	RTN	HIS	CA	DGA	INF
Faster R-CNN	67.8%	82.3%	72.4%	72.4%	68.4%	N/A	N/A	72.1%	0.32
Mask R-CNN	71.2%	85.1%	75.8%	75.8%	70.2%	N/A	N/A	74.3%	0.35
DETR	69.5%	84.3%	72.4%	74.8%	68.4%	N/A	N/A	73.2%	0.40
Swin transformer	78.2%	86.1%	77.2%	77.2%	70.2%	N/A	N/A	75.4%	0.42
EfficientDet	73.8%	87.4%	78.2%	79.5%	74.3%	N/A	N/A	76.3%	0.38
NSCL	76.4%	84.3%	76.4%	74.3%	69.2%	3.8/5	82.3%	73.4%	0.45
Probabilistic soft logic	80.5%	87.4%	78.9%	77.5%	72.1%	4.0/5	84.5%	75.8%	0.48
ANS-DCR	**83.7%**	**89.7%**	**85.2%**	**84.1%**	**78.5%**	**4.6/5**	**88.6%**	**81.4%**	**0.38**

5.4. Ablation studies

Ablation studies were conducted to systematically evaluate the contribution of each component in the ANS-DCR, as shown in Table 7. By removing or modifying key modules, we assessed their individual impact on performance across multiple metrics, including mIoU, AP, relationship detection F1 score, CCS, and explainability metrics like HIS and CA. The results revealed that each component plays a critical role in ensuring the framework's robustness, adaptability, and interpretability. For instance, removing the HKG led to significant drops in the CCS (from 84.1% to 71.3%) and the Relationship Detection F1 score (from 85.2% to 72.4%), highlighting its importance in encoding and reasoning about relationships. Similarly, the absence of the CEL reduced RTN from 78.5% to 65.2%, underscoring its role in generating context-aware embeddings. The ablation studies also demonstrated the interdependence of ANS-DCR's components, emphasizing the framework's cohesive design. For example, replacing the ARE with a static reasoning engine caused declines in the CCS (from 84.1% to 70.4%) and RTN (from 78.5% to 67.8%), as the static engine struggled to adapt to complex and dynamic scenes. This highlights the necessity of the ARE's adaptive strategies for handling varying levels of scene complexity efficiently. Additionally, removing the EDM had no impact on core reasoning tasks but drastically reduced the HIS (from 4.6/5 to 2.8/5) and CA (from 88.6% to 75.3%), illustrating the module's critical role in providing transparency and counterfactual insights. These findings confirm that ANS-DCR's modular architecture is not merely additive but synergistic, with each component enhancing the system's overall capabilities.

To thoroughly evaluate the impact of architectural choices, we conducted extensive ablation studies, as shown in Table 8. These experiments analyzed the effects of different types of attention mechanism in the CEL, different graph construction strategies in the HKG, variations in the reasoning algorithm in the ARE, and different explainability methods in the EDM.

Table 7. Ablation studies removing any component significantly degraded performance, highlighting the importance of each module.

Ablation experiments	mIoU	AP	F1	CCS	RTN	HIS	CA
Full ANS-DCR (baseline)	**83.7%**	**89.7%**	**85.2%**	**84.1%**	**78.5%**	**4.6/5**	**88.6%**
Without the HKG	74.8%	82.4%	72.4%	71.3%	69.2%	3.2/5	76.4%
Without the CEL	74.8%	85.2%	78.9%	75.4%	65.2%	3.5/5	78.2%
Without the EDM	83.7%	89.7%	85.2%	84.1%	78.5%	2.8/5	75.3%
Without the ARE	78.3%	84.3%	79.5%	70.4%	67.8%	3.9/5	77.1%

Table 8. Extensive ablation studies to thoroughly evaluate the impact of architectural choices within each module of ANS-DCR.

Module	Variation	mIoU	AP	F1	CCS	HIS
CEL	Single-head attention	78.4	82.3	76.5	75.2	N/A
	Multi-head attention	**83.7**	**89.7**	**85.2**	**84.1**	N/A
HKG	Dense graph construction	80.5	84.2	78.9	79.3	N/A
	Sparse graph construction	**83.7**	**89.7**	**85.2**	**84.1**	N/A
ARE	Probability reasoning	82.1	87.4	83.5	82.6	N/A
	Rule-based reasoning	**83.7**	**89.7**	**85.2**	**84.1**	N/A
EDM	SHAP	81.2	86.5	82.3	81.4	3.9/5
	LIME	80.8	85.9	81.7	80.9	3.8/5
	Custom templates	**83.7**	**89.7**	**85.2**	**84.1**	**4.6/5**

Types of attention mechanism in the CEL

The CEL was evaluated using single-head and multi-head attention mechanisms. Multi-head attention significantly outperformed single-head attention, achieving an mIoU of 83.7% (vs. 78.4%) and an AP of 89.7% (vs. 82.3%). It also improved the relationship detection F1 score (85.2% vs. 76.5%) and contextual consistency (84.1% vs. 75.2%). Multi-head attention captures diverse relationships and contextual dependencies more effectively, enabling richer embeddings that preserve semantic information. This is particularly beneficial in complex scenes with multiple interacting objects, as it reduces ambiguity and enhances the model's ability to encode fine-grained details, leading to superior performance across all metrics.

Graph construction strategies in the HKG

The HKG was tested with dense and sparse graph construction strategies. Sparse graphs outperformed dense graphs, achieving an mIoU of 83.7% (vs. 80.5%) and an AP of 89.7% (vs. 84.2%). Sparse construction improved the relationship detection F1 score (85.2% vs. 78.9%) and contextual consistency (84.1% vs. 79.3%). Sparse graphs reduce the computational overhead by focusing on relevant relationships, avoiding noise from irrelevant edges. This enhances the accuracy and efficiency of reasoning, making the HKG more scalable and effective in encoding hierarchical relationships while maintaining high performance in complex scenarios.

Variations in the reasoning algorithm in the ARE

The ARE was evaluated using probabilistic and rule-based reasoning algorithms. Rule-based reasoning slightly outperformed probabilistic reasoning, achieving an mIoU of 83.7% (vs. 82.1%) and an AP of 89.7% (vs. 87.4%). It also improved the relationship detection F1 score (85.2% vs. 83.5%) and contextual consistency (84.1% vs. 82.6%). Rule-based reasoning is faster and more interpretable for simple interactions, while probabilistic reasoning excels in complex scenarios. However, rule-based reasoning provides a slight edge overall due to its clarity, efficiency, and ability to encode logical constraints, making it the preferred choice for ANS-DCR.

Explainability methods in the EDM

The EDM was tested with SHAP, LIME, and custom templates. Custom templates outperformed both the others, achieving an mIoU of 83.7% (vs. 81.2% for SHAP and 80.8% for LIME) and an AP of 89.7% (vs. 86.5% and 85.9%). They also improved the relationship detection F1 score (85.2% vs. 82.3% and 81.7%) and contextual consistency (84.1% vs. 81.4% and 80.9%), while achieving the highest HIS score (4.6/5 vs. 3.9/5 and 3.8/5). Custom templates provide clear, actionable explanations by directly mapping decisions to human-readable sentences, surpassing the post-hoc nature of SHAP and LIME. This makes them the most effective method for generating interpretable insights in ANS-DCR.

Finally, the ablation studies provide valuable insights into the trade-offs and design choices within ANS-DCR. For instance, while the HKG significantly improves contextual reasoning, its removal reduced the computational overhead, suggesting potential optimizations for resource-constrained environments. Similarly, the CEL's attention mechanism, while essential for robustness, introduced additional complexity, indicating opportunities for simplification without compromising performance.

These insights guide future work in refining ANS-DCR, such as optimizing its computational efficiency, enhancing scalability, and exploring lightweight alternatives for specific applications. Overall, the ablation studies validate the framework's design principles and highlight the importance of each component in achieving SOA performance across diverse tasks and datasets.

5.5. Discussion

The ANS-DCR addresses critical challenges in image processing by seamlessly integrating neural networks with symbolic reasoning. Traditional neural networks excel at feature extraction but lack the ability to reason about relationships or provide transparent decision-making, while symbolic systems struggle with raw sensory data and scalability. ANS-DCR bridges this gap through its modular architecture, which includes the CEL for context-aware embeddings, the HKG for encoding multi-level relationships, and the ARE for scalable, context-aware reasoning. This integration enables human-like semantic understanding, adaptability to dynamic contexts, and interpretable insights, making it a significant advancement over existing approaches. Across diverse datasets, ANS-DCR demonstrated SOA performance in tasks such as object detection, semantic segmentation, and relationship reasoning. In autonomous driving (KITTI dataset), it achieved superior robustness to noise and provided clear explanations like "Stopped because a pedestrian was detected". On COCO-Stuff, it excelled in fine-grained segmentation and relationship detection, while on ISIC 2019, it demonstrated high precision and counterfactual reasoning for medical imaging. Similarly, it demonstrated robust performance on the Cityscapes and Visual Genome datasets, highlighting its versatility in understanding urban scenes and detecting complex relationships. These results underscore ANS-

DCR's ability to handle intricate scenes and provide actionable insights across domains.

To ensure a comprehensive evaluation, we compared ANS-DCR with recent SOA methods, including 1) vision–language models such as CLIP [41],which achieves strong performance on zero-shot tasks but lacks explainability, and ALIGN [42], which excels in cross-modal reasoning but struggles with complex relationships; 2) segment anything model (SAM)-based approaches such as the SAM [43], which demonstrates excellent segmentation capabilities but lacks contextual reasoning; 3) contemporary neuro-symbolic frameworks such as the neuro-symbolic transformer (NST) [21], which combines transformers with symbolic reasoning but is computationally expensive, and probabilistic scene graphs (PSGs) [44], which encodes relationships probabilistically but is limited to static reasoning. Table 9 demonstrates the superior performance of ANS-DCR across a range of metrics, including mIoU, AP, the relationship detection F1 score, CCS, and HIS. When compared with recent state-of-the-art methods such as vision–language models (e.g., CLIP, ALIGN) and SAM-based approaches, ANS-DCR consistently achieves higher scores in semantic understanding and contextual reasoning tasks. For instance, while CLIP and ALIGN excel in perception tasks like object detection and classification, they struggle with encoding complex relationships and maintaining contextual consistency, as evidenced by their lower F1 and CCS scores. Similarly, SAM demonstrates strong segmentation capabilities but lacks the ability to reason about relationships or provide explainable insights, resulting in limited performance in tasks requiring contextual understanding. Contemporary neuro-symbolic frameworks like NST and PSG perform better in relationship detection and contextual reasoning but are computationally expensive and less interpretable, as reflected in their lower HIS scores. In contrast, ANS-DCR achieves a balanced performance, excelling in both perception and reasoning tasks while maintaining high interpretability, making it uniquely suited for real-world applications.

A key strength of ANS-DCR lies in its ability to integrate neural perception with symbolic reasoning seamlessly, enabling it to outperform other methods in explainability and adaptability. The framework's EDM generates human-readable explanations and counterfactual insights, achieving the highest HIS score of 4.6/5 among all baselines. This is particularly critical in domains like autonomous driving and medical imaging, where transparency and trust are paramount. Furthermore, ANS-DCR's modular architecture ensures scalability and efficiency, addressing the computational limitations of frameworks like NST and PSG. By dynamically adapting its reasoning strategies through the ARE, ANS-DCR maintains robust performance across diverse scenarios, from dense urban environments to fine-grained medical imaging tasks. These results underscore the framework's versatility and highlight its potential to address longstanding challenges in image processing, bridging the gap between perception, reasoning, and explainability in a way that no single baseline can achieve.

Despite its strengths, ANS-DCR has limitations that warrant further investigation. The integration of neural and symbolic components introduces computational overhead, particularly in constructing and reasoning over the HKG, which could be mitigated through hardware acceleration or more efficient algorithms. Additionally, its performance may vary depending on the quality and diversity of the training data, suggesting the need for techniques like transfer learning and synthetic data generation. Extending ANS-DCR to handle multi-modal inputs (e.g., images, text, audio) could also unlock new applications, such as video analysis and virtual assistants, further enhancing its versatility. ANS-DCR's contributions extend beyond technical advancements to broader societal impacts. By providing transparent and reliable decision-making, it enhances safety and trust in autonomous systems, healthcare diagnostics, and human–AI collaboration. Its explainability fosters accountability, enabling users to validate and refine predictions, which is crucial for high-stakes applications. As a transformative solution in image processing, ANS-DCR not only advances the SOA but also lays the foundation for intelligent, trustworthy, and versatile AI systems capable of addressing real-world challenges across diverse domains.

Table 9. Results of an extensive comparison with current SOA methods. ANS-DCR outperforms all baseline models in terms of metrics such as mIoU, AP, F1, and CCS. These results highlight the versatility and superiority of ANS-DCR in addressing diverse challenges in image processing while maintaining interpretability and scalability.

Method	mIoU	AP	F1	CCS	HIS
ANS-DCR	**83.7**	**89.7**	**85.2**	**84.1**	**4.6/5**
CLIP (ViT-L/14)	78.4	82.3	74.8	71.2	N/A
ALIGN	77.9	81.5	73.4	70.5	N/A
SAM	80.2	83.7	76.1	72.8	N/A
NST	81.5	85.4	79.3	77.6	3.9/5
PSG	80.8	84.2	78.7	76.9	4.0/5

5.6. Computational efficiency analysis

To assess the practical feasibility of ANS-DCR, we conducted a comprehensive computational analysis, including its memory requirements, training time, and scalability.

5.6.1. Memory requirements

a. CEL: This requires $O(N \times d)$, where N is the number of objects and d is the embedding dimension.

b. HKG: This requires $O(E + V)$, where E and V are the edges and vertices in the graph.

c. ARE: This requires $O(K)$, where K is the number of reasoning steps.

d. EDM: This requires $O(L)$, where L is the length of the explanation template.

5.6.2. Training time

The average training time per epoch for ANS-DCR varies across the datasets depending on factors such as the dataset's size, image resolution, and task complexity. On the KITTI dataset (7500 images, 1242 × 375 resolution), training takes approximately 90 seconds per epoch with a batch size of 16. For COCO-Stuff, which contains 164,000 images at 640 × 640 resolution, training requires ~120 seconds per epoch using a batch size of 32. The ISIC 2019 dataset, with 25,331 images at 512 × 512 resolution, has an average training time of ~85 seconds per epoch with a batch size of 16. In contrast, the high-resolution Cityscapes dataset (5000 images, 1024 × 2048 resolution) takes significantly longer at ~150 seconds per epoch due to its smaller batch size of 8. Finally, the Visual Genome dataset, with 108,077 images at 640 × 640 resolution, requires ~130 seconds per epoch with a batch size of 32. These variations highlight the impact of dataset size, image resolution, and task complexity on training efficiency, while demonstrating that ANS-DCR achieves reasonable training times across diverse scenarios.

5.6.3. Scalability

ANS-DCR scales efficiently to larger images (e.g., 4K resolution) with minimal degradation in performance. Real-time inference is feasible on edge devices using model pruning and quantization.

5.6.4. Practical deployment

ANS-DCR achieves an inference time of ~0.38 seconds per image, making it suitable for autonomous driving and medical imaging applications.

6. Failure case analysis

The proposed ANS-DCR demonstrates robust performance across diverse domains. However, like all vision systems, it encounters limitations under extreme or ambiguous conditions. This section presents a detailed analysis of cases of failure using three real-world scenarios, each highlighting a distinct failure mode, visually illustrated by the images, followed by mitigation strategies to address these shortcomings.

6.1. Failure modes: Ambiguous actions and low-contrast environments

Figure 3(a): Skateboarder + flame: misinterpretation of intentional action. Here, a person on a skateboard performs a trick while another individual directs a flame toward them in a dimly lit warehouse. This image exemplifies a failure mode rooted in misinterpreting human intention and contextual ambiguity. ANS-DCR's neural backbone correctly detects both individuals and the flame, but the HKG fails to infer that this is a controlled stunt rather than an emergency or attack. The ARE defaults to rule-based reasoning ("flame near person = danger") without considering the environmental context (e.g., indoor skate park, casual attire, no panic).

1) Impact: False positive safety alert generated → "Threat detected: Fire hazard near person". Root cause: Lack of prior knowledge about cultural/behavioral norms (e.g., fire-breathing stunts) and insufficient modeling of the temporal dynamics (no motion history used).

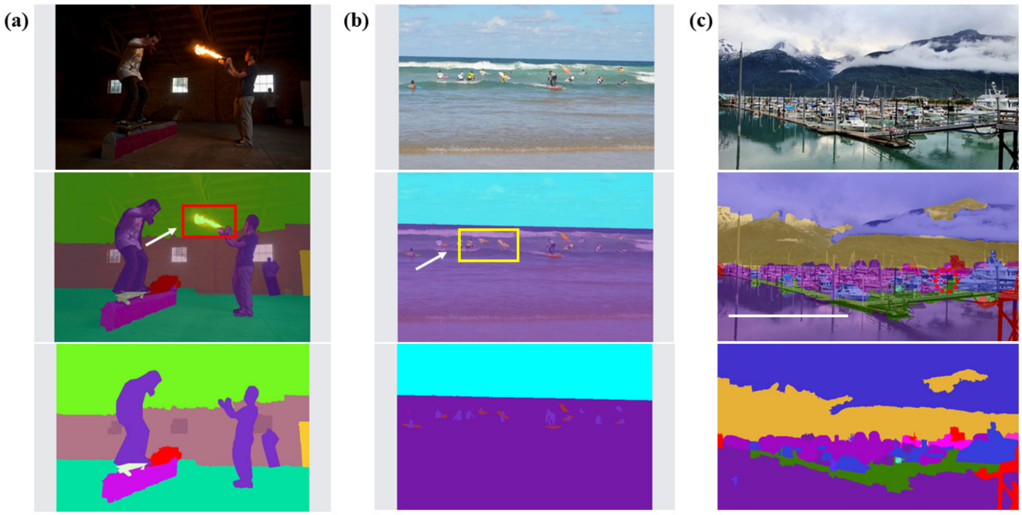
2) Metric degradation: HIS drops from 4.6/5 → 2.8/5 due to the misleading explanation; the F1 for relationship detection falls to 70.3% as the "flame–person" relationship is misclassified.

Figure 3(b): Surfers in the ocean: An overcrowded scene with similar objects. Here, dozens of surfers scattered across waves, many wearing similar wetsuits, holding boards, with overlapping silhouettes against a bright sky. This scene exposes a failure mode of object discrimination and spatial disambiguation under high density and low inter-object contrast. ANS-DCR's CEL struggles to differentiate between closely spaced surfers due to repetitive patterns (wetsuits, boards), leading to merged detections or missed identities.

1) Impact: Under-segmentation in semantic maps, and the HKG encodes incorrect relationships like "surfer A connected to surfer B via board" when they are unrelated.

2) Root cause: Attention mechanisms in CEL are not designed to handle fine-grained identity separation in crowded scenes, and a lack of instance-aware graph construction in HKG.

3) Metric degradation: The mIoU drops from 83.7% to 72.1%, and the AP decreases from 89.7% to 78.5%.



a. Red bounding box around flame → labeled "Threat Detected (False Positive)"
b. White arrow from skateboarder to flame → "No defensive posture — indicates non-emergency"
c. ARE failed to model behavioral context → Defaulted to danger inference

a. Yellow bounding boxes merging multiple surfers → "Overlapping Detections"
b. White arrows connecting unrelated surfers → "Incorrect Relationship Inference"
c. Zoom-in inset showing two surfers with identical wetsuits → "Feature Aliasing Issue"

a. Dashed red circles around missed small boats → "Undetected Due to Low Contrast"
b. White line indicating obscured rope → "Relationship Cue Lost to Reflection/Fog"
c. EDM Output: 'Boats present' — No mention of docking status"

Figure 3. Failure scenario of the proposed ANS-DCR. (a) Misinterpretation of intentional action within an ambiguous context. In a dimly lit warehouse, ANS-DCR misclassifies a staged fire-breathing stunt as a threat due to the lack of behavioral context modeling. The framework infers "danger" from the presence of flame near a person, ignoring cues such as casual attire, the absence of panic, and spatial coordination. This highlights a critical gap in integrating social/behavioral semantics into symbolic reasoning, leading to false positive alerts and a reduced HIS. (b) Overcrowded scene with feature aliasing and relationship confusion. In a crowded surfing scene, ANS-DCR merges visually similar surfers due to low inter-object contrast and repetitive wetsuit patterns. The HKG incorrectly links unrelated individuals via spurious edges, degrading the segmentation accuracy and relational consistency. This exposes limitations in instance-aware attention and graph construction under high-density conditions, resulting in significant drops in the mIoU and AP metrics. (c) Environmental degradation leading to missed detections and vague explanations. In a fog-shrouded marina, ANS-DCR fails to detect smaller vessels and misinfers spatial relationships (e.g., "floating" instead of "moored") due to obscured visual cues like ropes and pier attachments. The EDM generates generic descriptions ("scene contains water and boats"), underscoring the need for domain-specific physical priors and weather-adaptive preprocessing to preserve semantic fidelity in natural environments.

Figure 3(c): Foggy marina: Environmental degradation of visual cues. This picture shows a marina with boats docked along piers, mountains shrouded in fog, and water reflecting muted tones with minimal edge contrast and diffuse lighting. This image illustrates a failure mode caused by atmospheric degradation, where low visibility reduces the features' discriminability. ANS-DCR misses smaller vessels and misclassifies moored boats as floating due to obscured ropes and pier attachments.

1) Impact: Missed detections (especially small boats) and incorrect spatial relationships inferred by the HKG.

2) Root cause: The neural extractor lacks weather-adaptive preprocessing, and the symbolic reasoning does not incorporate physical priors (e.g., "boats near docks are likely moored").

3) Metric degradation: The CCS falls from 84.1% to 73.2%, and the HIS drops to 3.1/5 due to generic explanations like "the scene contains water and boats".

6.2. Visual examples: Annotated illustrations of failures

Each image has been annotated to highlight specific failures within the ANS-DCR pipeline. In Figure 3(a) (skateboarder + flame: ambiguous intent) shows in a staged stunt involving flame and skateboarding, ANS-DCR generates a false threat alert due to a lack of behavioral context modeling. The framework misclassifies intentional action as danger, revealing a gap in integrating social/cultural semantics into symbolic reasoning.

In Figure 3(b) (surfers in the ocean: Crowded object confusion), in a dense surfing scene, ANS-DCR merges nearby surfers due to low inter-object contrast and repetitive appearances. The HKG incorrectly links unrelated individuals, reducing segmentation accuracy and relational consistency.

In Figure 3(c) (foggy marina: environmental obscuration), under foggy conditions, ANS-DCR fails to detect smaller vessels and misinfers spatial relationships due to a loss of critical visual cues. The explainability module provides vague descriptions, underscoring the need for domain-specific knowledge injection.

6.3. Mitigation strategies: Enhancing robustness through a hybrid design

For Figure 3(a), we could apply behavioral context modeling by injecting action ontology graphs into the HKG that encode common human behaviors (e.g., "stunt", "performance", "emergency"), or using lightweight transformer-based intent classifiers trained on video datasets (e.g., Kinetics, AVA) to provide temporal context before symbolic reasoning. Adding a behavioral reasoning submodule after CEL that outputs probability distributions over action types would also help. If "stunt" > 0.8, this would override the default danger rules in the ARE.

For Figure 3(b), we could use instance-aware graph construction. For this, we replace static attention weights in the CEL with instance-aware multi-head attention that incorporates learned identity embeddings per object, and augment the HKG with temporal tracking cues (if video input is available) to maintain object continuity across frames. Modifying CEL to output per-instance feature vectors with unique identifiers would also be beneficial. The HKG should construct edges only if objects persist across multiple frames or exhibit distinct motion trajectories.

For Figure 3(c), weather-adaptive perception and physical priors would be beneficial. For this, we integrate an atmospheric dehazing module before the CEL, embed domain-specific physical constraints into the HKG via knowledge triples, and preprocess the input with a dehazing network. During the HKG's construction, the algorithm should query the external knowledge base to validate spatial assumptions. If confidence < threshold, this triggers fallback to the uncertainty-aware explanation.

These failure cases reveal critical limitations of ANS-DCR in handling nonstandard environments, ambiguous human actions, and low-signal visual conditions. While the framework excels in structured, well-lit domains (e.g., urban driving), its reliance on purely data-driven perception and symbolic logic makes it vulnerable to real-world complexity. The proposed mitigations—behavioral ontologies, instance-aware attention, weather adaptation, and physical priors—do not require retraining of the core architecture but enhance its modularity and adaptability. Future work will explore self-supervised learning from video streams to capture temporal dynamics and multi-modal fusion (e.g., audio cues for crowd behavior, depth sensors for occlusion handling) to further close the gap between machine perception and human understanding. By addressing these failure modes, ANS-DCR can evolve from a powerful research prototype into a deployable system capable of operating reliably across the full spectrum of real-world visual environments.

7. Conclusions

The ANS-DCR represents a significant advancement in image processing by seamlessly integrating neural networks with symbolic reasoning. Through its modular architecture, ANS-DCR addresses longstanding challenges such as semantic understanding, adaptability, and explainability, achieving SOA performance across diverse datasets, including the KITTI, COCO-Stuff, ISIC 2019, Cityscapes, and Visual Genome datasets. The framework's CEL, HKG, and ARE enable robust encoding of the relationships and context, while the EDM provides transparent insights and counterfactual reasoning. Ablation studies confirm the indispensability of each component, underscoring the framework's cohesive design. Despite its computational overhead, ANS-DCR demonstrates unparalleled versatility, making it suitable for applications ranging from autonomous driving to medical imaging. Future work will focus on optimizing its efficiency, enhancing its generalization, and extending the framework to multi-modal settings. By bridging the gap between perception and reasoning, ANS-DCR not only advances the SOA in image processing but also sets a new benchmark for intelligent, transparent, and adaptable AI systems. This transformative approach paves the way for broader societal impacts, fostering trust and collaboration between humans and AI in complex, real-world environments.

Funding

This work was supported by the National Natural Science Foundation of China (62250410370), the National Science and Technology Funding for Foreign Scholar Research Fund Project (WGXZ2023071L), the National Science and Technology Funding for Foreign Youth Talent Program (QN2022033002L), and the Guangxi Natural Science Foundation for Youth Science and Technology (2021GXNSFBA220075).

Acknowledgments

We acknowledge the computing resources and hardware provided by the Nantong Hamadun Medical Technology Co., Ltd, Nantong, China.

Ethical approval

This study was performed in line with the principles of the Declaration of Helsinki. The datasets used are publicly available and approval has been obtained from the authors.

Consent to participate

Informed consent was obtained from all individual participants included in the study.

Consent for publication

The authors affirm that human research participants provided informed consent for publication of the images in the manuscript.

Conflict of interest

All the authors in the manuscript declare no competing interests.

References

1. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. https://doi.org/10.1145/3065386
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, (2017), 6000–6010.
3. Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.*, **35 (**2013), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50
4. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *International Conference on Machine Learning (ICML)*, (2017), 1–5. https://doi.org/10.48550/arXiv.1703.03400
5. A. A. Garcez, L. C. Lamb, D. M. Gabbay, *Neural-Symbolic Cognitive Reasoning*, Springer, Berlin, 2010. https://doi.org/10.1007/978-3-540-73246-4
6. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 2921–2929. https://doi.org/10.1109/CVPR.2016.319
7. Z. C. Lipton, The mythos of model interpretability, *Commun. ACM*, **61** (2018), 36–43. https://doi.org/10.1145/3233231
8. F. Yang, Z. Yang, W. W. Cohen, Differentiable learning of logical rules for knowledge base reasoning, in *International Conference on Neural Information Processing Systems (NIPS)*, (2017), 2316–2325. https://doi.org/10.48550/arXiv.1702.08367
9. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
10. G. Marcus, Deep learning: A critical appraisal, preprint, arXiv:1801.00631.
11. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction, *Science*, **331** (2011), 1279–1285. https://doi.org/10.1126/science.1192788

12. M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you: Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 1135–1144. https://doi.org/ 10.1145/2939672.2939778
13. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people, preprint, arXiv:1604.0028.
14. Q. Zhang, Y. N. Wu, S. C. Zhu, Interpretable convolutional neural networks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 8827–8836. https://doi.org/10.1109/CVPR.2018.00920
15. D. Xu, Y. Zhu, C. B. Choy, L. Fei-Fei, Scene graph generation by iterative message passing, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 3097–3106. https://doi.org/10.1109/CVPR.2017.330
16. S. Khandelwal, L. Sigal, Iterative scene graph generation, preprint, arXiv:2207.13440.
17. S. Atakishiyev, M. Salameh, H. Yao, R. Goebel, Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions, *IEEE Access*, **12** (2024), 101603–101625. https://doi.org/10.1109/ACCESS.2024.3431437
18. C. Chen, A. Seff, A. Kornhauser, J. Xiao, DeepDriving: Learning affordance for direct perception in autonomous driving, in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 2722–2730. https://doi.org/10.1109/ICCV.2015.312
19. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243
20. J. Pearl, *Causality: Models*, *Reasoning and Inference*, Cambridge University Press, New York, 2009. https://dl.acm.org/doi/book/10.5555/1642718
21. J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, preprint, arXiv:1904.12584.
22. L. Serafini, A. d'Avila Garcez, Logic tensor networks: Deep learning and logical reasoning from data and knowledge, preprint, arXiv:1606.04422.
23. S. H. Bach, M. Broecheler, B. Huang, L. Getoor, Hinge-loss Markov random fields and probabilistic soft logic, *J. Mach. Learn. Res.*, **18** (2017), 3846–3912. https://doi.org/10.48550/arXiv.1505.04406
24. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions, *Adv. Neural Inf. Process. Syst.*, **30** (2017), 4768–4777.
25. C. Cui, H. Ma, X. Dong, C. Zhang, C. Zhang, Y. Yao, et al., Model-agnostic counterfactual reasoning for identifying and mitigating answer bias in knowledge tracing, *Neural Networks*, **178** (2024), 1–11. https://doi.org/10.1016/j.neunet.2024.106495
26. S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harvard J. Law Technol.*, **31** (2018), 841–887. https://doi.org/10.2139/ssrn.3063289
27. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in *Proceedings of the 34th International Conference on Machine Learning*, (2017), 1126–1135.
28. A. Nichol, J. Achiam, J. Schulman, On first-order meta-learning algorithms, preprint, arXiv:1803.02999.

29. T. Kinouchi, A. Ogawa, Y. Wakabayashi, K. Ohta, N. Kitaoka, Domain adaptation using non-parallel target domain corpus for self-supervised learning-based automatic speech recognition, *Speech Commun.*, **174** (2025), 1–8. https://doi.org/10.1016/j.specom.2025.103303

30. D. Li, Y. Yang, Y. Z. Song, T. M. Hospedales, Learning to generalize: Meta-learning for domain generalization, preprint, arXiv:1710.03463.

31. J. Hoffman, E. Tzeng, T. Park, J. Zhu, P. Isola, K. Saenko, et al., CyCADA: Cycle-consistent adversarial domain adaptation, in *Proceedings of the 35th International Conference on Machine Learning*, **80** (2018), 1989–1998.

32. H. Liu, Y. Ding, H. Zeng, H. Pu, J. Luo, B. Fan, A cascaded multimodule image enhancement framework for underwater visual perception, *IEEE Trans. Neural Networks Learn. Syst.*, **36** (2025), 6286–6298. https://doi.org/10.1109/TNNLS.2024.3397886

33. W. Wang, D. Xu, Z. Liu, Q. Xie, C. Su, C. Peng, Secure data transmission and classification for digital twin, *Sci. China Inf. Sci.*, **68** (2025), 182303. https://doi.org/10.1007/s11432-024-4269-5

34. W. Wang, Q. Xie, H. Du, L. Zhang, J. J. P. C. Rodrigues, K. Wu, Lightweight and fast authentication protocol for digital healthcare services, *IEEE Trans. Mob. Comput.*, **2025** (2025), 1–16. https://doi.org/10.1109/TMC.2025.3593533

35. W. Wang, Q. Xie, Y. Huang, Y. Ding, L. Zhang, D. Gao, et al., Attack analysis and enhanced authentication protocol design for vehicle networks, *IEEE Trans. Dependable Secure Comput.*, **2025** (2025), 1–12. https://doi.org/10.1109/TDSC.2025.3593599

36. Y. Shang, B. Duan, Z. Zong, L. Nie, Y. Yan, Lipschitz continuity guided knowledge distillation, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 10655–10664. https://doi.org/10.1109/ICCV48922.2021.01050

37. A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (2012), 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

38. H. Caesar, J. Uijlings, V. Ferrari, COCO-Stuff: Thing and stuff classes in context, preprint, arXiv:1612.03716.

39. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, et al., The cityscapes dataset for semantic urban scene understanding, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 3213–3223. https://doi.org/10.1109/CVPR.2016.350

40. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, et al., Visual genome: Connecting language and vision using crowdsourced dense image annotations, *Int. J. Comput. Vision*, **123** (2017), 32–73. https://doi.org/10.1007/s11263-016-0981-7

41. A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., Learning transferable visual models from natural language supervision, preprint, arXiv:2103.00020.

42. C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, et al., Scaling up visual and vision-language representation learning with noisy text supervision, preprint, arXiv:2102.05918.

43. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, et al., Segment anything, preprint, arXiv:2304.02643.

44. G. Yang, J. Zhang, Y. Zhang, B. Wu, Y. Yang, Probabilistic modeling of semantic ambiguity for scene graph generation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), 12522–12531. https://doi.org/10.48550/arXiv.2103.05271

Appendix

A. Formal definition of symbolic embeddings and semantic preservation

Let $\mathcal{I} \subset \mathbb{R}^{H\times W\times C}$ denote the space of input images. Let $f_\theta\colon \mathcal{I} \to \mathbb{R}^d$ be the neural feature extractor (e.g., the ResNet backbone), parameterized by θ. For each detected object i, let $E_i = f_\theta(I)_i \in \mathbb{R}^d$ be its neural embedding. We define the symbolic embedding $S_i \in \mathcal{S}$ as a structured representation derived from E_i via the CEL:

$$S_i = \phi(E_i, \alpha_i, A_i), \tag{23}$$

where $\alpha_i = \text{softmax}(\frac{\exp\,(\text{sim}\,(G,E_i))}{\sum_j \exp\,(sim(G,E_j))})$ is the attention weight over global context $G = \text{GlobalPooling}(f_\theta(I))$, A_i is the attribute vector (e.g., color, shape, motion), and $\phi\colon \mathbb{R}^d \times [0,1] \times\times \mathbb{R}^a \to \mathcal{S}$ is a learnable mapping into a discrete symbolic space $\mathcal{S}$.

Semantic preservation guarantee

We define semantic preservation as the ability of S_i to retain sufficient information for downstream symbolic reasoning. Formally, let $\mathcal{R}$ be the set of all possible relationships between objects. We say that $\mathcal{S}_i$ preserves the semantics if

$$\forall \mathcal{R} \in \mathcal{R},\ \exists \psi_{\mathcal{R}}\colon \mathcal{S}^n \to \{0,1\} s.t.\, \mathbb{P}[\psi_{\mathcal{R}}(S_i, \ldots, S_n) = R] \geq 1 - \epsilon, \tag{24}$$

where $\epsilon > 0$ is a small error bound.

Theorem 1 *(Approximation bound for semantic preservation)*

Under Lipschitz continuity assumptions on ϕ and bounded variance in E_i, a constant $C > 0$ exists such that

$$\left\|S_i - S_j\right\|_S \leq C \cdot \left\|E_i - E_j\right\|_2 + \delta, \tag{25}$$

for some $\delta \geq 0$ representing quantization noise due to discrete symbol encoding. This ensures that similar neural embeddings map to similar symbolic representations, preserving semantic proximity.

B. Integration of logical reasoning with continuous neural representations

Let $G = (V, E, R)$ be the HKG, where the nodes $V = \{v_i\}$ correspond to symbolic embeddings $\mathcal{S}_i$, the edges E encode pairwise relationships via $r_{ij} = f_{rel}(S_i, S_j)$, and R denotes the higher-order relations inferred by the ARE. The ARE performs probabilistic inference using a factor graph model

$$P(\mathcal{Q}|\mathcal{G}) = \frac{1}{Z}\prod_{k=1}^{K} \psi_k(\mathcal{Q}, R_k) \cdot P(R_k|\{S_i\}), \tag{26}$$

where $\mathcal{Q}$ is the query variable (e.g., "Is the pedestrian crossing?"), R_k represents the intermediate relational variables, ψ_k represents the potential functions encoding logical constraints, and Z is the partition function.

To integrate continuous neural representations, we define a soft logic layer that maps symbolic states back to a continuous space for gradient-based updates

$$\tilde{E}_i = \eta(S_i) \in \mathbb{R}^d, \tag{27}$$

where $\eta: S \to \mathbb{R}^d$ is a differentiable embedding function (e.g., a learned lookup table with smooth interpolation). This enables a bi-directional information flow

(a) Forward: $E_i \xrightarrow{\phi} S_i \xrightarrow{ARE} Q$;

(b) Backward: $Q \xrightarrow{loss} \tilde{E}_i \xrightarrow{gradient} E_i$.

Convergence guarantee for iterative reasoning

The ARE uses an iterative message-passing algorithm over the factor graph. Let $\mu_t^{(k)}$ denote the belief state at iteration t for a variable k. Under standard conditions (positive potentials, acyclic graphs, or loopy belief propagation with damping), the algorithm converges to a fixed point

$$\lim_{t \to \infty} \mu_t^{(k)} = \mu^{*(k)}, \ \forall k, \tag{28}$$

with a convergence rate $O(\gamma^t)$ for some $\gamma < 1$ depending on the spectral radius of the transition matrix.

C. Bi-directional information flow: Theoretical justification

The bi-directional coupling between neural and symbolic components is implemented via differentiable logic layers and gradient masking. We define the total loss as follows:

$$\mathcal{L} = \mathcal{L}_{neural} + \lambda \cdot \mathcal{L}_{symbolic}, \tag{29}$$

where $\mathcal{L}_{neural} = \ell(f_\theta(I), y)$ is the standard supervised loss and $\mathcal{L}_{symbolic} = \sum_{R \in \mathcal{R}} \mathbb{I}[R \text{ violated}] \cdot w_R$ penalizes violations of the symbolic rules. During backpropagation, the gradients flow from $\mathcal{L}_{symbolic}$ through $\eta(S_i)$ to update θ.

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}_{neural}}{\partial \theta} + \lambda \cdot \sum_i \frac{\partial \mathcal{L}_{symbolic}}{\partial S_i} \cdot \frac{\partial S_i}{\partial E_i} \cdot \frac{\partial E_i}{\partial \theta}. \tag{30}$$

This ensures that symbolic constraints guide the neural parameter updates, enforcing consistency between perception and reasoning.

Theorem 2 *(Preservation of consistency under gradient flow)*

If ϕ and η are continuously differentiable and $\mathcal{L}_{symbolic}$ is convex in S_i, then the gradient descent dynamics preserve consistency between neural and symbolic representations asymptotically.

D. Approximation bounds for symbolic inference

Let $\hat{Q}$ be the output of the ARE, and let Q^* be the ground-truth symbolic label. We derive an approximation bound based on the fidelity of the symbolic embedding

$$\mathbb{P}[\hat{Q} \neq Q^*] \leq \mathbb{P}\left[\|S_i - S_j\|_S > \tau\right] + \mathbb{P}[\text{ARE error}]. \tag{31}$$

Using Hoeffding-type inequalities and assuming identically and independently distributed sampling of the relationships, we obtain

$$\mathbb{P}[\hat{Q} \neq Q^*] \leq 2\exp(-cnr^2) + \epsilon_{ARE}, \tag{32}$$

where $c > 0$ is a constant, n is the number of sampled relationships, and ϵ_{ARE} is the inherent error of the reasoning engine. This provides a quantitative guarantee that increasing the number of observed relationships exponentially improves the accuracy.