

Article

Not peer-reviewed version

Detection and Recognition of Visual Geons Based on Specific Object of Interest Imaging Technology

Wu Yonghao , Liu Minyi , [Li Jun](#) *

Posted Date: 25 March 2025

doi: 10.20944/preprints202503.1897.v1

Keywords: object of interest imaging; visual geons; visual attention; deep learning



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Detection and Recognition of Visual Geons Based on Specific Object of Interest Imaging Technology

Wu Yonghao, Liu Minyi and Li Jun *

School of Electronic Science and Engineering, South China Normal University, Guangzhou, Guangdong 510631, China

* Correspondence: lizhjcd@scnu.edu.cn

Abstract: In various fields including visual processing, computer graphics, neuroscience, and biological sciences, geons are widely recognized as fundamental units of complex shapes. Their importance has been broadly acknowledged. However, accurately identifying and extracting these geons remains a challenge. This study integrates theories from signal processing, computer graphics, neuroscience, and biological sciences, utilizing "specific object of interest imaging" and neural networks to describe mathematical operators, in order to reveal the essence of visual geons. Experiments validate the core hypothesis of geon theory, namely that geons are foundational components for the visual system to recognize complex objects. Through training, neural networks are capable of identifying distinct basic geons and, based on this foundation, performing target recognition in more complex scenarios. This effectively confirms the existence of geons and their critical role in visual recognition, providing new tools and theoretical foundations for related research fields.

Keywords: object of interest imaging; visual geons; visual attention; deep learning

1. Introduction

The generation and parsing of complex geometric structures generally adhere to hierarchical construction principles, whereby finite core topological elements form higher-order visual representations through combinatorial rules. The academic community widely acknowledges that the fundamental units capable of representing all visual graphic expressions are defined as "geons"(geometric ions), also termed geometric primitives or simply primitives. Since the establishment of Marr's computational theory of vision, theoretical advancements in the formal representation of geons have continuously progressed across computer graphics, cognitive neuroscience, and biological sciences, sufficiently demonstrating the pivotal role of geon theory in explainable artificial intelligence and bio-inspired computing.

1.1. RBC (Recognition-by-Components) Framework Based on Marr

In the 1980s, Biederman and others proposed the Recognition-by-Components (RBC) theory [1,2]. This theory systematically explained the conceptual framework of geons, suggesting that the human visual system recognizes objects through a limited set of basic geometric shapes (no more than 36) and their spatial relationships. These geons possess volumetric properties. Around the same time, Marr's visual processing theory proposed that when processing images, the visual system first conducts a preliminary analysis of pixels, progressively extracting low-level features such as edges, lines, and brightness changes from lower to higher levels, forming "primitives" [3–15]. Marr's multi-level visual representation model constructs complex geometric shapes by detecting and combining basic geons, thereby enabling comprehensive scene recognition. This model emphasizes optimizing information transmission at each level under specific rules, generating raw sketches as the foundation for advanced visual tasks such as stereo matching, depth estimation, and motion detection. Marr proposed the application of mathematical methods and computational models to extract these fundamental primitives from images, simulating the principles of the human visual system in three-

dimensional object recognition, and established the "computational-algorithmic-implementational" three-level model during his research:

1. Computational level: Primitives as the minimal units of visual representation;
2. Algorithmic level: Deep learning simulates hierarchical feature extraction;
3. Implementation level: Weight optimization in neural networks (θ, ξ) corresponds to synaptic plasticity.

This hierarchical and modular processing structure bridges low-level perceptual information with high-level visual tasks, facilitating object segmentation, motion analysis, and the cognition of complex visual scenes, thus effectively constructing high-level representations of real-world scenes. Later, Tomaso Poggio and others regarded "basic primitives" as the theoretical basis for visual processing and object recognition. Their work on "I-Theory" explored how the visual cortex identifies shapes and objects through the progressive abstraction of image features, alongside his research on invariant visual representations [16–23]. Minsky and his co-authors discussed pattern recognition and the concept of decomposing complex shapes into simple elements in Perceptrons, emphasizing the visual components required for artificial intelligence to successfully recognize objects. His "Society of Mind" theory further developed a hierarchical processing framework with geons as cognitive units [24–30].

The main idea of the RBC framework can be summarized as object recognition being achievable by decomposing objects into basic components and identifying their relationships, a process referred to as "componential analysis." This core idea involves three key processes: edge extraction, component recognition, and matching. RBC has three principles:

1. Constructing complex objects: By combining and arranging different basic geons, various complex visual objects can be created, offering high flexibility.
2. Providing recognition cues: Basic geons not only form the shape and structure of objects but also provide key cues for object recognition. When an observer sees an object, they first identify the basic geons of the object, and then recognize the object based on the combination and arrangement of these elements.
3. Supporting rapid recognition: Due to the limited number and types of basic geons, and the relatively fixed ways they can be combined, the visual system can quickly recognize objects, making this process highly automated.

However, scholars of the RBC framework pointed out in their retrospective reflections that this theoretical framework has the following core deficiencies. Firstly, the representation system of visual geons suffers from conceptual confusion, with ambiguous definitions regarding key issues like feature detection mechanisms and topological structure analysis. Secondly, it fails to clarify the dynamic weight distribution between data-driven and concept-driven processing. Thirdly, and most critically, there is a lack of a unified computational framework for multimodal perceptual integration, which restricts the hierarchical integration of perceptual information. This absence of a system-level architecture ultimately constrains the overall emergence of the visual perception representation system. This critical reflection reveals the fundamental limitations of early computational vision theories in terms of neural interpretability and cognitive integration mechanisms.

1.2. Deep Learning-Based CV Scholars

Since the 21st century, the research focus of visual workers has shifted from traditional theoretical exploration to the development of practical deep learning technologies. The HPNet model proposed by Yan S et al. achieves end-to-end high-precision basic shape segmentation by integrating pixel-level semantic segmentation and basic shape segmentation [31]. For 3D shape generation, the 3D-PRNN developed by Zou C et al. employs a coding-decoding architecture combined with recurrent neural networks, effectively addressing the limitations of traditional methods in diversity and structural detail capture [32]. Lin Y et al. combined synthetic data with deep neural networks to achieve object structure representation based on limited parameters, successfully applied to robotic grasping tasks [33]. In the field of graphic generation, the neural point graphic model proposed by Aliev K A innovatively uses

point clouds to represent geometric shapes, significantly reducing the computational cost of processing complex geometries while generating high-quality graphic results [34]. The H3DNet developed by Zhang Z et al. achieves precise parsing and object reconstruction of LiDAR point cloud data through mixed geometric geons representation and a multi-branch network structure [35]. Xia S et al. provided a systematic review of the detection, normalization, and application methods of geometric geons across multiple disciplines, providing significant references for related research [36]. Li et al. consider "basic primitives" to be a core concept in the sketch segmentation process, where the model can segment and automatically label the identified basic geons [37]. In action recognition, Wang H et al. innovatively treat skeletal joints as geometric basic geons, achieving richer action feature extraction through multi-level information representation learning [38]. Hilbig A et al. suggest that using signed distance fields to encode basic geons can effectively enhance the generalization ability of 3D CNNs in geometric shape processing [39]. For 3D point cloud data processing, Romanengo C et al. established a benchmark testing framework based on SHREC, providing a standardized platform for evaluating geometric geon recognition algorithms [40]. The deep geometric segmentation method proposed by Li D et al. achieves precise fitting of basic geometric geons in 3D point clouds [41]. In the field of image parsing, the "DeepPrimitive" method developed by Huang J et al. enables recognition and segmentation of image components through basic geons decomposition [42]. Geometry C S et al. used a neural parser to build a framework for automatic parsing and reconstruction of complex 3D shapes based on basic geons [43]. These studies collectively advance the development of computer vision and graphics technologies based on basic geons, providing new solutions for complex scene understanding and shape reconstruction.

In the evolution of the theory of visual geons, although deep learning-based methods have made significant progress, the fundamental issues they face still need to be explored in depth. Marr emphasized in his hierarchical processing framework that the neglect of the organizational principles of biological visual systems in computer vision research may lead researchers to focus on pseudo-problems arising from limitations in sensors, hardware, or computational capacity, rather than on the essential issues of visual cognition. Although this paradigm shift in research provides many important insights in the short term, it may limit the in-depth development of the discipline in the long run. This warning is significantly corroborated in contemporary image segmentation research—despite long-term study, its theory and practice still face significant challenges, stemming from two fundamental issues. On one hand, accurately defining the target of segmentation in images and the physical world itself poses inherent difficulties. The philosophical definition of object-hood leads to a semantic definition dilemma regarding "entity-background" separation. On the other hand, there is an essential difference at the neural computational level between the inductive biases of existing algorithms and the adaptive representation mechanisms of biological visual systems. Current segmentation methods often rely on specific prior assumptions and constraints, which may not adequately reflect the flexibility and adaptability of the human visual system.

1.3. Psychology, Neuroscience, and Other Fields

In psychology and neuroscience, significant progress has also been made in the study of geons. Treisman's Feature Integration Theory (FIT) centers on visual geon and systematically explains the hierarchical processing mechanism of the human visual system, from parallel geon extraction to sequential attentional scanning. The theory posits that visual cognition initiates with rapid parallel extraction of geons, followed by spatial attention binding discrete geons into coherent objects. In visual cortical areas (e.g., V1-V4), contextual suppression and feature extraction are believed to be mediated by neurons' selective responses. The V1 area is responsible for extracting elementary low-level features, whereas the higher-tier V4 area integrates these low-level features into complex visual information through feature binding mechanisms [44–55]. This discovery not only establishes the foundational role of geons in visual processing, providing a seminal model for understanding visual attention and geon recognition, but also furnishes neuroscientific justification for the specific object of interest imaging methodology subsequently introduced in this study. Itti's team constructed a visual saliency

computation model based on the feature integration theory, proposing a saliency model through three innovative dimensions: center-surround differential algorithms, multi-scale feature extraction, cross-modal normalization, and re-entrant suppression mechanisms. They believe that the saliency map constructed by basic geons can predict shifts in attentional focus, and that different feature channels have independent representational characteristics in the primary visual cortex, providing a computational framework for feedforward visual processing [56–63]. Scholars from South China Normal University have made systematic breakthroughs in the field of visual geon research. Research indicates that visual geons, as the foundational units of visual information encoding, play a core role in spatial information integration, early visual processing, and higher cognitive functions. Specifically, Hang Y et al. define visual geons as the foundational units by which the visual system encodes and processes visual stimuli, arguing that they play a core role in spatial information integration and early visual processing, particularly related to the perception of object size [64]. Yuxuan C et al. point out that visual geons are the building blocks for the brain’s processing of complex visual information; by decomposing visual stimuli and integrating these geons, the brain can achieve higher cognitive functions such as object recognition, spatial localization, and motion perception [65]. Zhang W et al. indicate that visual geons are fundamental elements in visual processing that influence people’s cognition and emotional responses to complex landscapes [66]. Wen X et al. found that as the complexity of visual tasks increases, the brain’s functional networks dynamically reorganize, particularly in regions involved in geon processing, showing tighter connectivity [67]. Experiments by Zhou LF et al. reveal that visual geons have advantages in both conscious and unconscious face perception; even in an unconscious state, geons can still aid in recognizing and distinguishing facial features, laying the groundwork for processing complex visual information [68]. Research by Zhang Y et al. shows that during the planning and execution of convergent movements, the visual system dynamically adjusts its focus and decoding of geons, thereby optimizing the perception of object size. This process involves the adjustment of activity in the visual cortex areas V1 to V4, which is significant for distance judgment and object detail recognition [69]. Zhu H et al. believe that the recognition and processing of visual geons are key to the emergence of stereoscopic depth perception; these geons serve as primary inputs for signal transmission, providing necessary input for higher-level neural mechanisms [70].

Table 1. Comparison of Theoretical Frameworks for Geon Representation.

Theoretical Framework	Definition of Geons	Dynamic Adaptability	Noise Robustness	Improvements in This Paper
RBC (Biederman)	Fixed 36 geometric shapes	Low	Not involved	Adaptive learning of geons
I-Theory	Gradual abstraction of features	Medium	Partial	Incorporation of sparse coding
DeepPrimitive	Data-driven action segmentation	High	Data-dependent	Cross-modal bio-inspired

The aforementioned studies have amply demonstrated the significance of geons across multiple disciplinary domains. Nevertheless, these studies exhibit marked limitations, particularly in experimental validation of theoretical models, necessitating the establishment of novel mathematical frameworks that integrate computational tractability with cognitive plausibility. Furthermore, numerous studies predominantly rely on hypothetical deduction and computational simulation, lacking systematic empirical data support. Current segmentation methods in computer vision (CV) deep

learning often depend on specific assumptions, demonstrating inadequate universal adaptability to multi-scenario and multi-task requirements.

2. Materials and Methods

Treisman's Feature Integration Theory (FIT) demonstrates the biological plausibility of contextual suppression and geon extraction, providing theoretical foundations for the feature binding mechanism in perceptual systems. This feature binding mechanism reveals that the human visual system can isolate and extract target geons within complex backgrounds, a process that not only enhances perceptual efficiency but also offers theoretical support for contextual suppression and object representation. Biederman's Recognition by Components (RBC) theory posits that geons, as independent visual units, must be distinctly segregated within complex backgrounds to ensure accurate object recognition. These theories have informed the conceptual framework of this research. Building upon the biological plausibility of contextual suppression and theoretical grounding of primitive segregation, this study incorporates attention mechanisms to implement a target-specific imaging methodology termed "specific object of interest imaging" proposed by Li et al [71,72]. This method aims to remove irrelevant backgrounds from images. By creating specific dataset labels, the neural network learns only the information of specific targets, ultimately achieving complete imaging of the target. This method optimizes the use of an end-to-end neural network, allowing it to learn from different pairs of training data. Each data pair must contain a panoramic image M_h and a label image N_h of a specific target, where $h = 1, 2, \dots$. The training process is similar to an optimization process and can be expressed as:

$$\{\xi, \theta\} = \operatorname{argmin}_{\xi, \theta} \frac{1}{H} \sum_{h=1}^H \|N_h - \tilde{N}_h\| \quad (1)$$

$$\tilde{N}_h = P_{\theta}(O_{\xi}(M_h)) \quad (2)$$

The training image M_h is processed through the compression network O_{ξ} and the reconstruction network P_{θ} to obtain the target image N_h , where ξ and θ are the weights of the compression and reconstruction networks respectively. $\|\cdot\|$ defines the loss function evaluating the approximation error between N_h and $\tilde{N}_h$. This method provides a novel approach for the field of salient object detection, while also demonstrating broad applicability across other computer vision application scenarios. By extracting key features of target objects and comparing them with hypothesized geons, this method enables rapid validation of theoretical models' rationality and stability. The object imaging-based geon theory enhances the robustness and adaptability of object representation, effectively integrating biologically inspired principles with computational models.

The "specific object of interest imaging" methodology is now implemented to experimentally realize primitive recognition and detection. Given a photographic input containing N hypothesized primitives, each object is formally represented as:

$$k_i, i = 0, 1, \dots, N \quad (3)$$

For the graphical objects of the entire scene, it can be represented as:

$$k = \sum_{i=0}^{N-1} k_i \quad (4)$$

Each graphical object k_i can be represented as a linear combination in an M -dimensional sparse space and projected onto this subspace. The sparse representation is achieved using the sparse basis ψ and sparse coefficients x within the sparse space, formulated as:

$$k_i = \sum_{j=1}^M \psi_{ij} x_{ij} \quad (5)$$

The panoramic image k can then be formally expressed as:

$$k = \sum_{i=0}^{N-1} k_i = \sum_{i=0}^{N-1} \sum_{j=1}^M \psi_{ij} x_{ij} \quad (6)$$

A simple network is employed to train on sample images and ground truth of target k' to capture its complete feature profile. After training, the neural network's parameters encapsulate the feature information of target k' , termed the feature filter Φ . Φ applies sparse transformation to panoramic image k , yielding estimated salient objects formulated as:

$$\Phi k = \sum_{i=0}^{N-1} \sum_{j=1}^M \Phi \psi_{ij} x_{ij} = \sum_{j=1}^M \psi_{0j} x_{0j} = \hat{k}_0 \quad (7)$$

Applying the feature filter Φ to other scene images allows for filtering and ultimately recognizing the assumed basic geons.

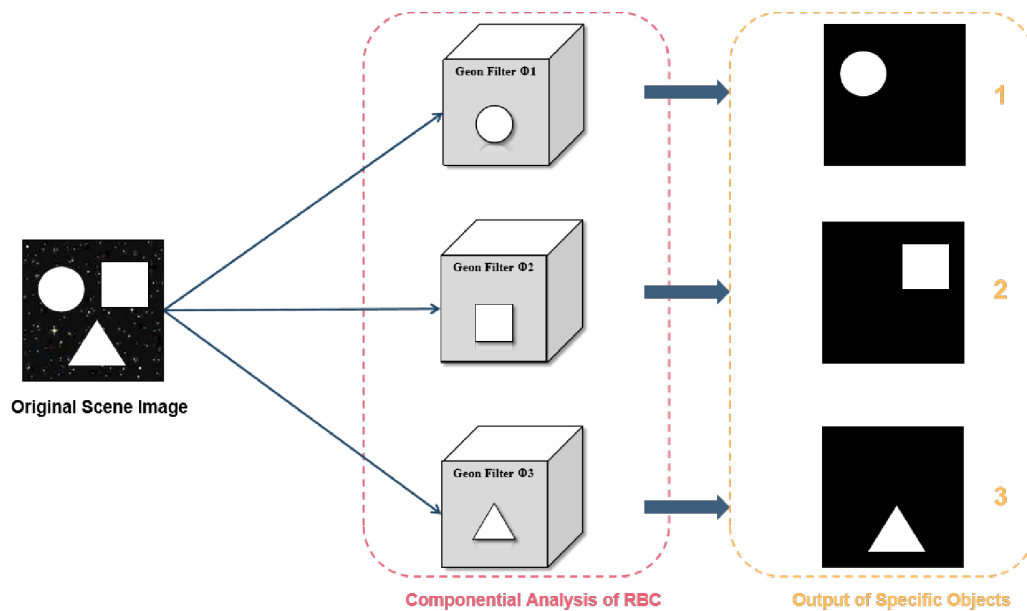


Figure 1. Technical Schematic of Object of Interest Imaging Methodology.

Considering th

at the mathematical operator of "geons" is difficult to express directly with mathematical functions, this study uses a simple universal neural network to describe this operator. By setting different input-output data pairs to train, optimize, and approximate this mathematical operator. The resulting neural network will be regarded as the desired "geons." U-Net is selected as the foundational network architecture [73]. U-Net exhibits unique advantages in multi-scale feature fusion and processing complex spatial relationships. Preliminary experiments in our laboratory demonstrated that compared to alternative architectures (e.g., MLP [74] and KAN [75]), U-Net's skip connections effectively integrate hierarchical features, emulating the layered processing mechanism of the human visual system. Skip connections enable direct utilization of encoder-stage low-level features during decoding, facilitating complementary integration of detailed low-level features with semantic high-level features. This mechanism achieves multi-scale feature fusion and cross-layer information transfer, proving critical for processing complex primitive topologies and geometric reconstruction. Comparatively, Multilayer Perceptron (MLP) lacks spatial information capture capability due to its fully-connected nature; Kernel Attention Network (KAN) shows limitations in multi-scale feature representation despite its attention mechanism focus; Half-U-Net's unilateral architecture undermines decomposition-recomposition capacity by underutilizing cross-layer connectivity. The symmetric encoder-decoder architecture of

U-Net aligns perfectly with the decomposition-recomposition primitive theory: The encoder pathway comprises four downsampling stages that progressively extract high-level semantic features while synchronously processing primitive characteristics from pixel-level details to global structures through hierarchical feature extraction, corresponding to primitive decomposition. The decoder pathway achieves precise localization via four upsampling stages, employing cross-layer fusion strategies to integrate low-level details with high-level semantics, maintaining spatial accuracy during geometric reconstruction - this pathway mirrors the recomposition process that restores spatial relationships and geometric configurations, establishing U-Net as the optimal choice for this task.

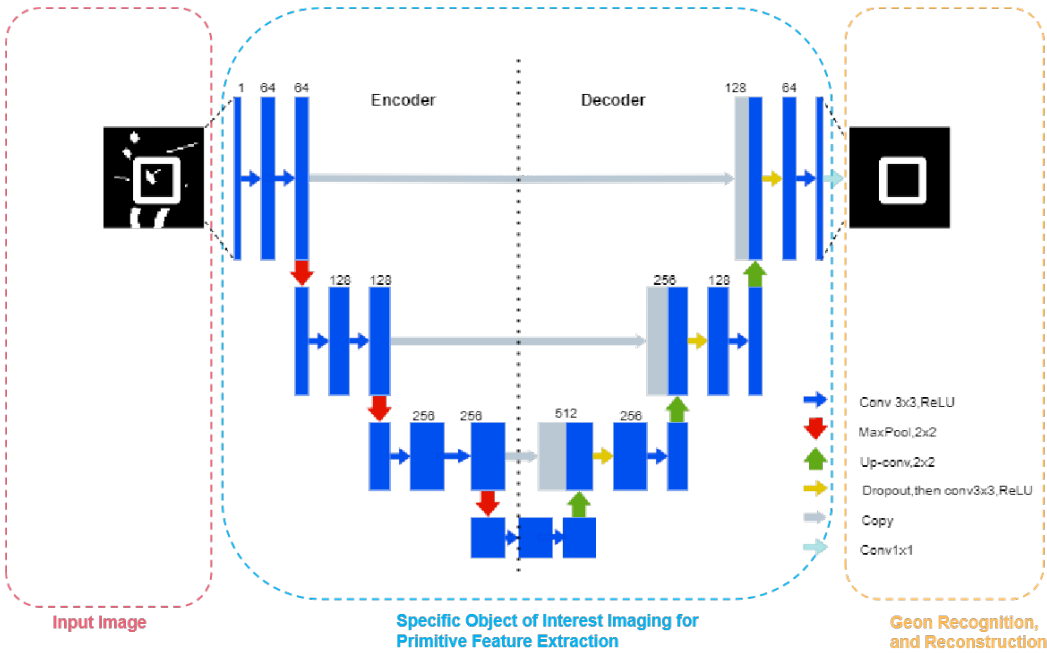


Figure 2. U-Net Architecture Schematic.

Corresponding to Marr’s tri-level framework, this study also constructs a "computational-algorithmic-implementational" three-level paradigm:

1. Computational level: Consistent with RBC theory, primitives serve as minimal units of visual representation.
2. Algorithmic level: Deep learning simulates hierarchical feature extraction, where U-Net’s encoder→decoder architecture corresponds to human V1→IT cortical pathways.
3. Implementation level: Neural network weight optimization (θ , ξ) mirrors synaptic plasticity, partially validating Zhang et al.’s visual cortical activity modulation [69].

Itti’s saliency model, grounded in biological visual attentional mechanisms, achieves salient region extraction through feature contrast computation. Compared to conventional fixed filters, this study introduces a dynamic feature filter (ϕ) design that exhibits theoretical congruence with Itti’s saliency framework. Distinct from the static nature of fixed filters, ϕ adaptively extracts saliency features across scales via end-to-end learning of geon characteristics, enabling precise target geon capture with concurrent background suppression, thereby resolving the "feature detection ambiguity" inherent in the RBC framework. Furthermore, our theoretical framework incorporates biological visual noise suppression mechanisms [56–63], providing explanatory power for model performance variance under different noise types, thus achieving closer alignment with Itti’s saliency principles. Experimental comparisons between Gaussian and Salt-and-Pepper noise conditions revealed superior robustness under Gaussian noise, attributable to its spectral continuity better approximating natural environmental interference patterns - a characteristic congruent with biological visual perception. Subsequent experimental validation demonstrates that our U-Net-based object imaging model effectively emulates

V1 cortical contrast adaptation mechanisms through cross-channel normalization, thereby enhancing noise suppression capabilities.

3. Experiments and Results

3.1. Datasheet

Building upon Biederman's 36-category geometric taxonomy and Zhou et al.'s findings on non-conscious geon perception [68], this study selects triangle, circle, and square as target geon to ensure broader applicability. A custom-built geon dataset was constructed for algorithmic validation, comprising three geometric categories with random perturbations, generating 1,100 synthetic images as the base dataset, with 1,000 for the training set and 100 for the validation set.

A multi-level evaluation framework is established following the principle of cognitive interpretability, constructing test sets hierarchically for different validation targets: Single geon with geometric disturbances, multiple geons with geometric disturbances, and multiple target geons with geometric disturbances serve as the basic validation sets, with 100 samples for each type; Two types of typical noise disturbances—additive Gaussian noise and salt-and-pepper noise—are applied to each of the three types of samples. Ultimately, this results in 9 types of test sets, each containing 100 images, totaling 900 test samples. The test sets establish a hierarchical evaluation scenario that includes single geon recognition, multiple geon separation, multi-target task detection, and noise disturbance reconstruction of multiple tasks, and corresponding test sets are used based on different task requirements.

3.2. Evaluation Indicators

This study employs the following three classic quantitative metrics for a systematic evaluation of algorithm performance.

1. SSIM [76]

Based on the multi-channel characteristics of the human visual system, its perceptual consistency metric is defined as:

$$SSIM(X, Y) = \frac{(2\mu_x\mu_y + C_1)(2\delta_{xy} + C_2)}{(\mu_x^2\mu_y^2 + C_1)(\delta_x^2\delta_y^2 + C_2)} \quad (8)$$

By comparing brightness, contrast, and structural similarity, visual perceptual consistency is quantified within the range of [0, 1].

2. PSNR [77]

As a frequency domain fidelity benchmark, its calculation model is:

$$PSNR = 20\log_{10}\left(\frac{MAX_i}{\sqrt{MSE}}\right) \quad (9)$$

Here, MAX_i represents the maximum pixel value of the image (e.g., 255 for an 8-bit image). This metric objectively reflects image fidelity by measuring the ratio of the maximum signal power to the distortion power.

3. MSE [78]

Defined as a convergence verification metric for the optimization process:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (10)$$

PSNR provides a global distortion benchmark, SSIM characterizes perceptual consistency, and MSE ensures the verification of algorithm convergence, together forming a complementary evaluation system. The experimental section will combine quantitative analysis with visual comparisons to ensure the completeness of the evaluation conclusions.

3.3. Experimental Setup Implementation Details

The deep learning hardware configuration employed in this study consists of an AMD Ryzen 7 6800H central processing unit (CPU) and an NVIDIA GeForce RTX 3050 Ti Laptop graphics processing unit (GPU). To ensure objective methodology comparison, strict adherence to the network architecture selected in the original study was maintained. The training strategy utilized Mean Squared Error (MSE) loss as the objective function. The training duration was adaptively configured 60 epochs based on geon recognition complexity, with a batch size of 4. The Adam optimizer was implemented with initial learning rates set 0.002, determined by geon recognition performance characteristics. Input specifications standardize all images to 64×64 pixel grayscale format with unified resolution.

3.4. Results

3.4.1. Fundamental Detection.

The training process employing the aforementioned methodology yielded diverse geon imaging recognition outcomes, with representative results partially displayed below.

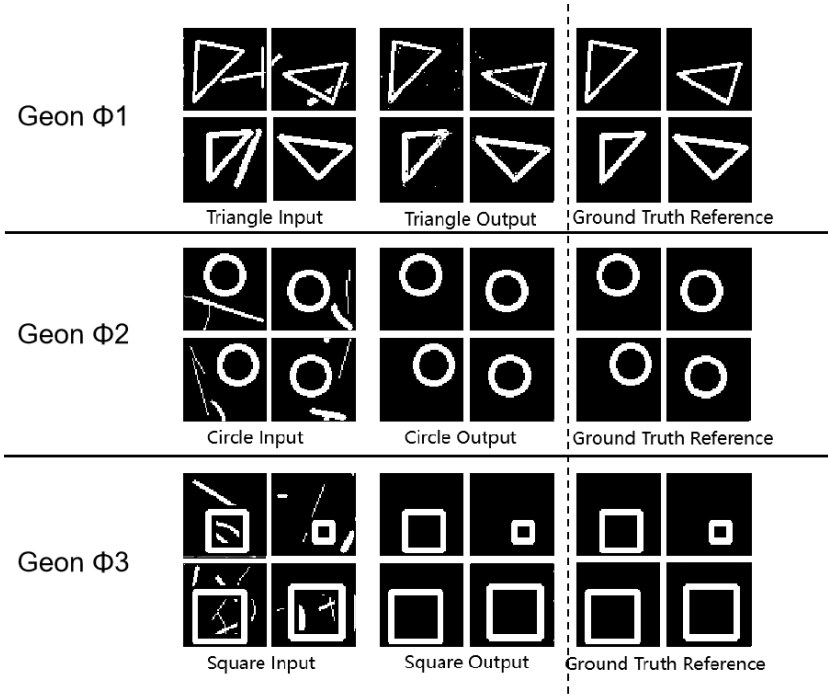


Figure 3. Experimental Results Diagram.

Table 2. Result Metrics.

Geon	SSIM	PSNR	MSE
Triangle	0.93	59.14	0.10
Circle	0.99	54.64	0.23
Square	0.98	58.40	0.11

Quantitative experimental evaluation demonstrates that the method developed in this study exhibits high precision and stability in graphic geon detection tasks. The structural similarity index (SSIM) values for all geons exceeded 0.93, with circular shapes achieving an SSIM of 0.99, indicating the method’s exceptional capability in reconstructing structural features of geons, particularly circular elements. The low peak signal-to-noise ratio (PSNR) values suggest minimal noise generation during geon reconstruction.

These tripartite metrics collectively validate the high precision of our geometric geon extraction approach. The experimental findings resonate with Marr’s geon theory at the neural-computational

level, where the feature extraction and combinatorial capacity of basic geon enable efficient visual object recognition.

3.4.2. Geon Separation Detection

This study extended the investigation to include geon separation experiments. The experimental results demonstrating separation efficacy across diverse geon types are selectively presented below.

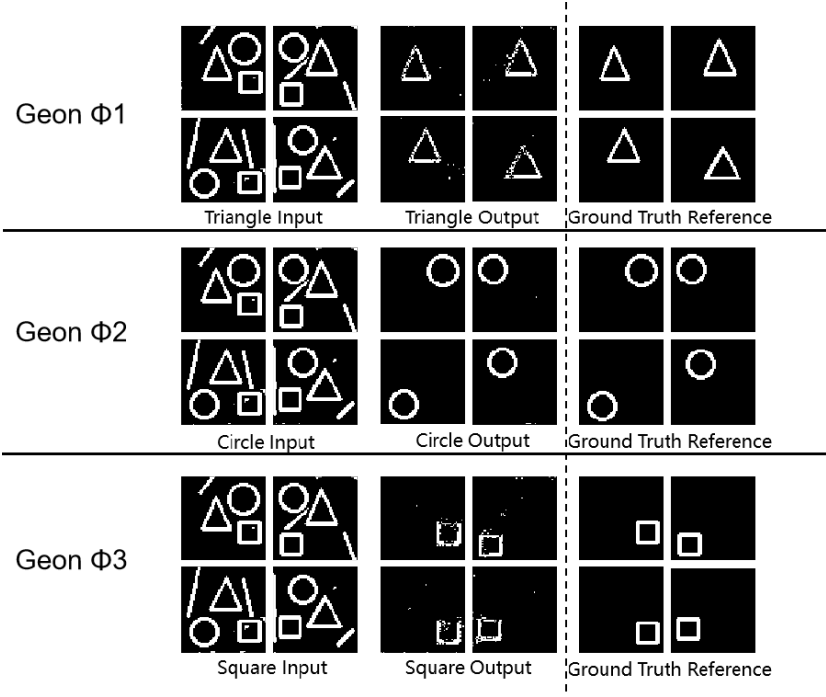


Figure 4. Experimental Results Diagram.

Table 3. Result Metrics.

Geon	SSIM	PSNR	MSE
Triangle	0.88	62.34	0.05
Circle	0.99	60.05	0.06
Square	0.80	58.88	0.10

Experimental data analysis demonstrates that the proposed method maintains effective separation and extraction of geometric geons in complex scenarios, showcasing superior separation accuracy. Particularly in circular geon separation experiments, achieving an SSIM of 0.99 and MSE of 0.06 indicates exceptional performance in circular element isolation, with complete structural preservation and effective noise suppression. The model also attains high fidelity and low error rates in triangular and square geon separation, showing only minimal feature detail degradation. These consolidated results confirm the method’s capability to efficiently dissect and extract structural features of fundamental geons from multi-geon composite images. This capability is most pronounced in circular separation tasks, where near-perfect accuracy demonstrates robust circular detection and feature extraction. Furthermore, our method exhibits substantial separation efficacy for other geons. Aligning with the core proposition of Marr’s geon theory, these experimental findings validate its fundamental tenet: visual systems can effectively recognize complex objects through deconstruction and feature isolation of basic geons.

3.4.3. Multi-geon Detection

To rigorously validate the network’s geon extraction and representational capacity while assessing generalization capabilities, further testing was conducted on its performance in multi-target scenarios. Selected experimental outcomes from this investigation are presented below.

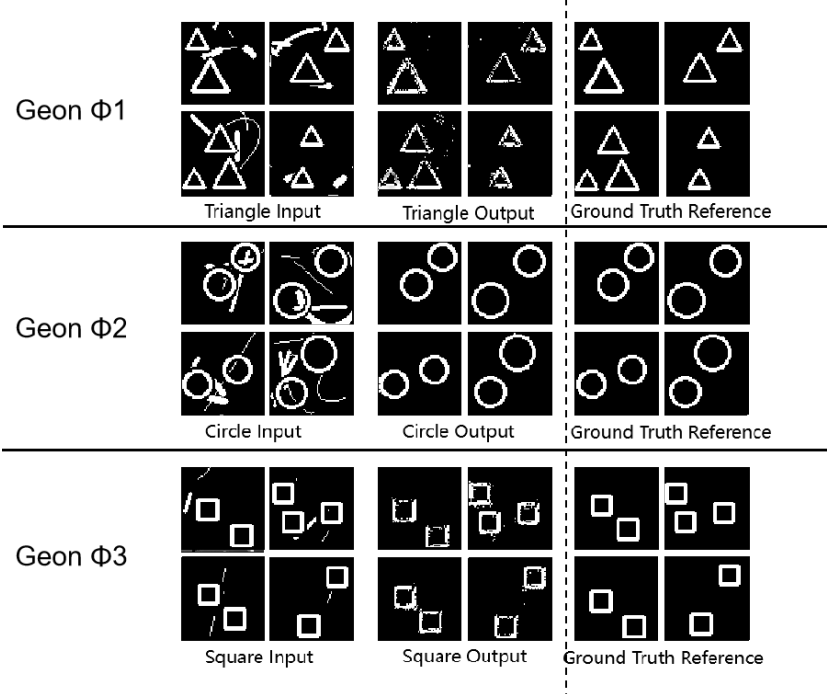


Figure 5. Experimental Results Diagram

Table 4. Result Metrics.

Geon	SSIM	PSNR	MSE
Triangle	0.79	46.17	2.64
Circle	0.98	48.58	1.35
Square	0.77	51.55	0.61

Experimental results demonstrate that the proposed methodology maintains effective geon extraction in multi-object scenarios, exhibiting robust feature separation and representation capabilities. Circular geons retain optimal performance, with the model accurately extracting structural features under multi-object conditions while maintaining low noise levels. This validates the method’s superior feature extraction and segmentation capacity for high-regularity geons like circles, demonstrating exceptional generalization performance. For triangular and square geons, while performance metrics show relative decline compared to circles, the model still achieves precise contour extraction with maintained structural integrity. This phenomenon aligns with Treisman’s attentional bottleneck theory [44–55], confirming theoretical predictions regarding feature integration limitations.

3.4.4. Robustness Evaluation Under Gaussian Noise Conditions

To validate the stability of the proposed method under noise interference, this study conducted systematic experiments on geons under Gaussian and salt-and-pepper noise conditions to evaluate robustness. For triangle and circular detection tasks (both single and multi-target configurations) which demonstrated superior performance in preliminary tests, Gaussian noise with $\sigma=30\text{--}35$ was introduced to the test sets. Square detection experiments employed $\sigma=17$ Gaussian noise, while geon separation tests utilized identical noise level ($\sigma=17$). Selected experimental results under these noise conditions are presented below:

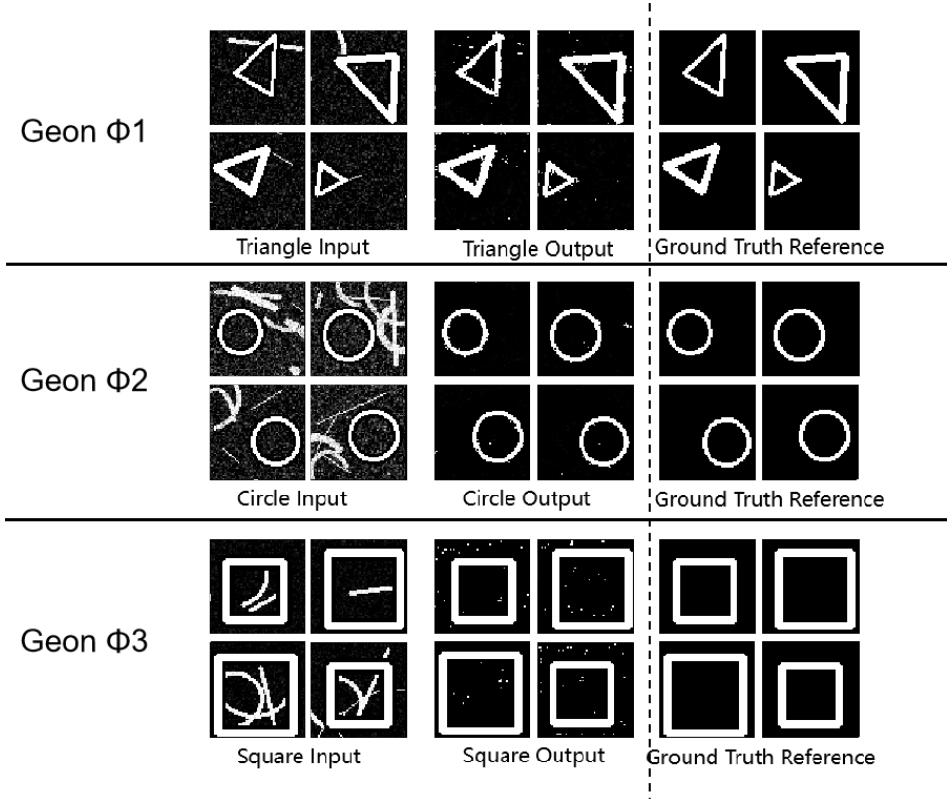


Figure 6. Experimental Results Diagram for Basic Detection Tasks.

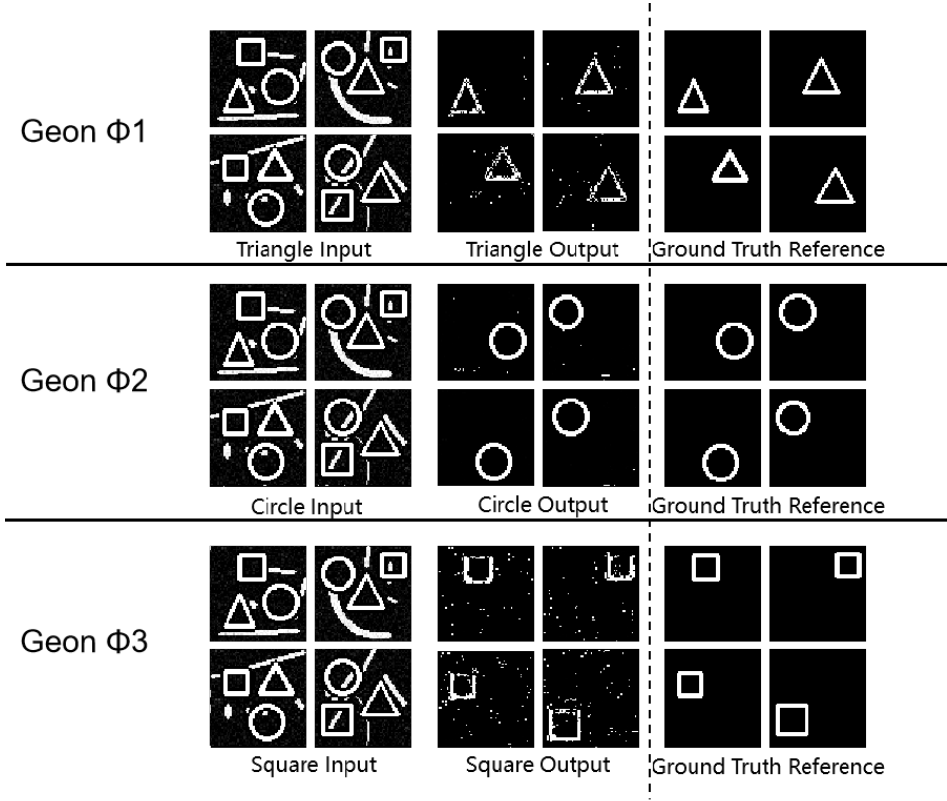


Figure 7. Experimental Results Diagram for Geon Separation.

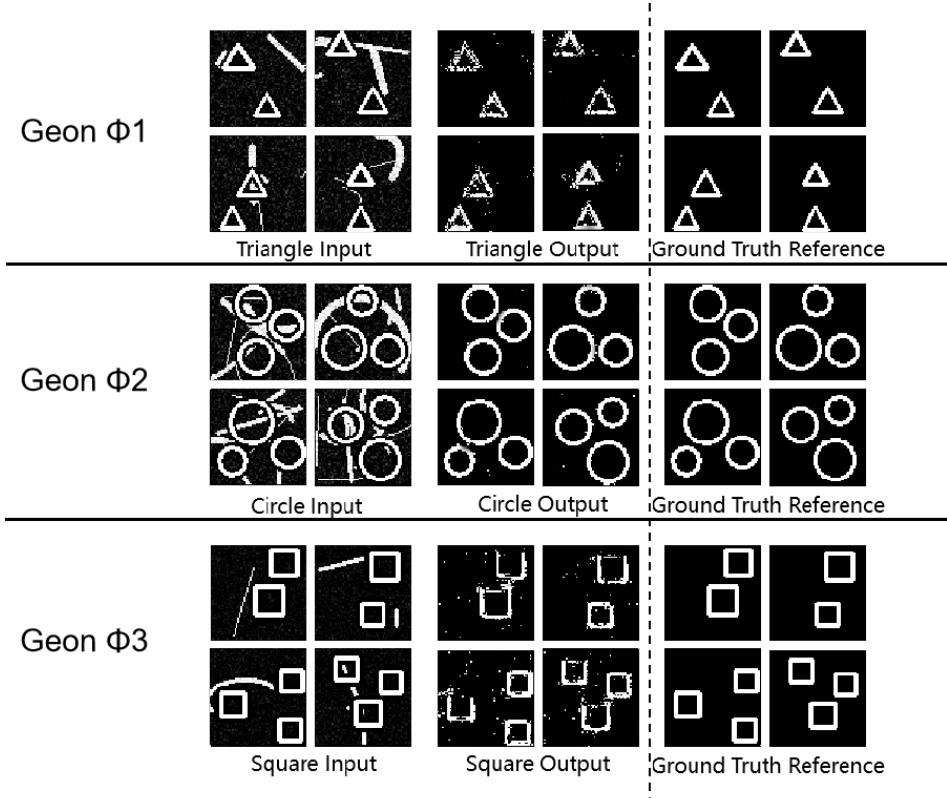


Figure 8. Experimental Results Diagram for Multi-Target Detection.

Integrated experimental analysis reveals that the proposed methodology achieves effective segregation between noise signals and geon structural information, with the encoder’s high-dimensional feature space demonstrating low sensitivity to noise distribution. Remarkably, the circular contour maintains an SSIM of 0.93 even under high noise levels, evidencing geon operators’ strong capture capacity for fundamental geons. This validates the superiority of the shape-encoding approach in noise resilience.

3.4.5. Robustness Evaluation under Salt-and-Pepper Noise Conditions (SPN)

Test sets for three geon basic detection tasks (demonstrating superior baseline performance) and circular multi-target detection were augmented with salt-and-pepper noise at density 0.05, while other datasets incorporated lower-density noise (0.02). Selected experimental outcomes under these configured noise scenarios are presented below:

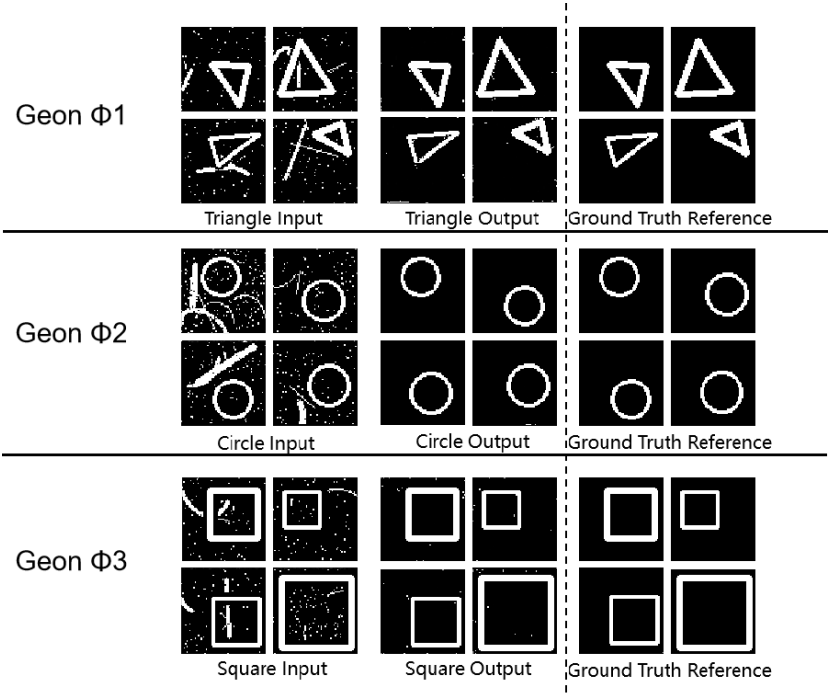


Figure 9. Experimental Results Diagram for Basic Detection Tasks.

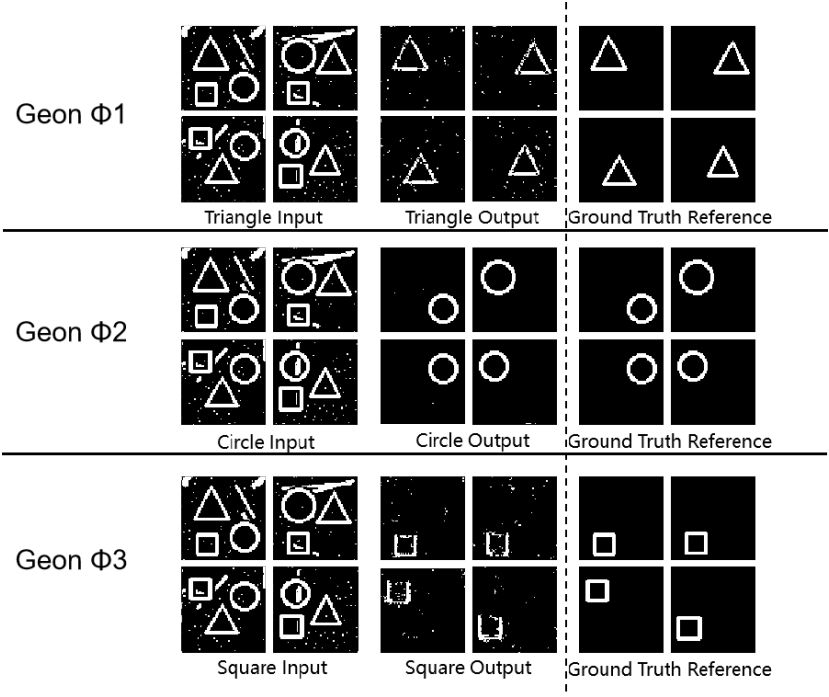


Figure 10. Experimental Results Diagram for Geon Separation.

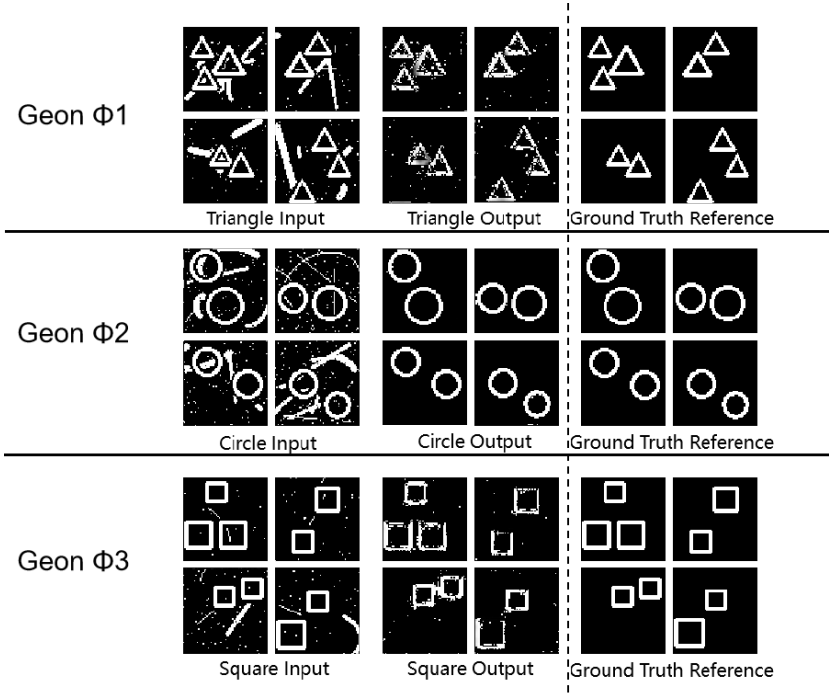


Figure 11. Experimental Results Diagram for Multi-Target Detection.

Experimental results demonstrate sustained robustness under multiplicative noise interference, with circular detection maintaining strong resilience at elevated noise densities. This advantage stems from the geometric regularity of circular geons - their inherent symmetry and smooth boundary characteristics mitigate noise impact on structural integrity.

The proposed method exhibits exceptional consistency across geon recognition and separation tasks, demonstrating remarkable generalization capabilities and noise immunity in multi-target scenarios. These results strongly align with Marr’s primal sketch theory, which posits that visual systems prioritize extraction of geometrically salient geons over pixel-wise matching. Through skip connections and hierarchical feature extraction, the geon operator effectively integrates local-global information flow, achieving particularly stable performance in detecting regular geons like circles. The framework demonstrates high-precision recognition/separation, cross-scene generalizability, and noise-resilient processing. This computational paradigm offers new pathways for constructing bio-inspired visual systems that emulate cortical information processing hierarchies.

4. Conclusions

Current analyses demonstrate that primitives hold extensive application prospects and significant research value across computer graphics, vision science, and artificial intelligence. Geon recognition proves critical not only for comprehending complex objects but also for elucidating the operational mechanisms of visual systems. While historical approaches like Marr’s methodology relied on predefined geon categories, this study centers on fundamental visual geons, establishing a multi-disciplinary convergence of psychological neuroscience, deep learning, and signal processing theories. We reveal how visual systems construct understanding of complex objects and scenes through hierarchical analysis and composition of basic geons. Through the "specific object of interest imaging" methodology and neural network-based mathematical operator characterization, we systematically investigate the existence and functionality of these visual building blocks. Experimental results synergize with theoretical principles to validate geon theory’s pivotal role in visual recognition, ultimately unveiling the ontological essence of geons. Robustness evaluations under Gaussian and Salt-and-Pepper noise conditions substantiate the theoretical framework’s superiority in noise-resistant geon recognition and detection. With advancing deep learning technologies, geon extraction methodologies will

undergo continuous refinement. Future work proposes integrating attention modules to simulate feature binding mechanisms, exploring Open-Set Geon Learning paradigms, and leveraging generative architectures (e.g., Diffusion Models) to emulate visual system plasticity.

References

1. Biederman, I. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing* **1985**, 32(1), 29–73.
2. Biederman, I. Recognition-by-components: a theory of human image understanding. *Psychological Review* **1987**, 94(2), 115.
3. Marr, D., Hildreth, E. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **1980**, 207(1167), 187–217.
4. Marr, D., Poggio, T. From understanding computation to understanding neural circuitry. **1976**.
5. Marr, D. Visual information processing: The structure and creation of visual representations. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **1980**, 290(1038), 199–218.
6. Marr, D. Early processing of visual information. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **1976**, 275(942), 483–519.
7. Marr, D. A theory for cerebral neocortex. *Proceedings of the Royal Society of London. Series B. Biological sciences* **1970**, 176(1043), 161–234.
8. Marr, D. Representing visual information. **1977**.
9. Marr, D., Thach, W. T. A theory of cerebellar cortex. In *From the Retina to the Neocortex: Selected Papers of David Marr* (1991), 11–50.
10. Marr, D. Analysis of occluding contour. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **1977**, 197(1129), 441–475.
11. Marr, D., Nishihara, H. K. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **1978**, 200(1140), 269–294.
12. Marr, D., Poggio, T. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **1979**, 204(1156), 301–328.
13. Marr, D., Ullman, S. Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **1981**, 211(1183), 151–180.
14. Marr, D. Artificial intelligence—a personal view. *Artificial Intelligence* **1977**, 9(1), 37–48.
15. Marr, D., Vaina, L. Representation and recognition of the movements of shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **1982**, 214(1197), 501–524.
16. Torre, V., Poggio, T. A. On edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1986**, 2, 147–163.
17. Bertero, M., Poggio, T. A., Torre, V. Ill-posed problems in early vision. *Proceedings of the IEEE* **1988**, 76(8), 869–889. .
18. Yuille, A. L., Poggio, T. A. Scaling theorems for zero crossings. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1986**, 1, 15–25.
19. Brunelli, R., Poggio, T. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1993**, 15(10), 1042–1052.
20. Sung, K. K., Poggio, T. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1998**, 20(1), 39–51.
21. Marr, D., Poggio, T. A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **1979**, 204(1156), 301–328.
22. Papageorgiou, C., Poggio, T. A trainable system for object detection. *International Journal of Computer Vision* **2000**, 38, 15–33.
23. Riesenhuber, M., Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neuroscience* **1999**, 2(11), 1019–1025.
24. Bhagyashri, P. T., Rupesh, T., Priyanka, K. K., et al. Marvin Minsky: The Visionary Behind the Confocal Microscope and the Father of Artificial Intelligence. *Cureus* **2024**, 16(9).
25. Minsky, M. L. Why people think computers can't. *AI Magazine* **1982**, 3(4), 3–3.
26. Minsky, M. A framework for representing knowledge. (1974-6-1)
27. Minsky, M., Papert, S. (1969) Marvin Minsky and Seymour Papert, Perceptrons. Cambridge, MA: MIT Press, Introduction, pp. 1–20, and p. 73 (figure 5.1). 1988.
28. Minsky, M. Commonsense-based interfaces. *Communications of the ACM* **2000**, 43(8), 66–73.

29. Minsky, M., Papert, S. An introduction to computational geometry. *Cambridge Tracts*, **1969**, 479(480), 104.
30. Minsky, M. Decentralized minds. *Behavioral and Brain Sciences* **1980**, 3(3), 439–440.
31. Yan, S., Yang, Z., Ma, C., Huang, H., Vouga, E., Huang, Q. HPNet: Deep Primitive Segmentation Using Hybrid Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 2753–2762.
32. Zou, C., Yumer, E., Yang, J., et al. 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), 900–909.
33. Lin, Y., Tang, C., Chu, F. J., et al. Using synthetic data and deep networks to recognize primitive shapes for object grasping. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), IEEE, 10494–10501.
34. Aliev, K. A., Sevastopolsky, A., Kolos, M., et al. Neural point-based graphics. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII* (2020), Springer International Publishing, 696–712.
35. Zhang, Z., Sun, B., Yang, H., et al. H3dnet: 3d object detection using hybrid geometric primitives. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII* (2020), Springer International Publishing, 311–329.
36. Xia, S., Chen, D., Wang, R., et al. Geometric primitives in LiDAR point clouds: A review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2020**, 13, 685–707.
37. Li, L., Fu, H., Tai, C. L. Fast sketch segmentation and labeling with deep learning. *IEEE Computer Graphics and Applications* **2018**, 39(2), 38–51.
38. Wang, H., Wang, L. Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. *IEEE Transactions on Image Processing* **2018**, 27(9), 4382–4394.
39. Hilbig, A., Vogt, L., Holtzhausen, S., et al. Enhancing three-dimensional convolutional neural network-based geometric feature recognition for adaptive additive manufacturing: A signed distance field data approach. *Journal of Computational Design and Engineering* **2023**, 10(3), 992–1009.
40. Romanengo, C., Raffo, A., Biasotti, S., et al. SHREC 2022: Fitting and recognition of simple geometric primitives on point clouds. *Computers & Graphics* **2022**, 107, 32–49.
41. Li, D., Feng, C. Primitive fitting using deep geometric segmentation. In *Proceedings of the International Symposium on Automation and Robotics in Construction* (2019), 36, 780–787.
42. Huang, J., Gao, J., Ganapathi-Subramanian, V., et al. DeepPrimitive: Image decomposition by layered primitive detection. *Computational Visual Media* **2018**, 4, 385–397.
43. Geometry, C. S. Neural Shape Parsers for Constructive Solid Geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, 44(5).
44. Treisman, A. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing* **1985**, 31(2), 156–177.
45. Treisman, A., Gormican, S. Feature analysis in early vision: evidence from search asymmetries. *Psychological Review* **1988**, 95(1), 15.
46. Treisman, A. Features and objects in visual processing. *Scientific American* **1986**, 255(5), 114B–125.
47. Treisman, A. Perceptual grouping and attention in visual search for features and for objects. *Journal of Experimental Psychology: Human Perception and Performance* **1982**, 8(2), 194.
48. Treisman, A. How the deployment of attention determines what we see. In *Progress in Psychological Science around the World. Volume 1: Neural, Cognitive and Developmental Issues* (2013), Psychology Press, 245–277.
49. Treisman, A. Binocular rivalry and stereoscopic depth perception. *Quarterly Journal of Experimental Psychology* **1962**, 14(1), 23–37.
50. Treisman, A. Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **1998**, 353(1373), 1295–1306.
51. Treisman, A. Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology Section A* **1988**, 40(2), 201–237.
52. Treisman, A., DeSchepper, B. Object tokens, attention, and visual memory. (1996).
53. Treisman, A., Paterson, R. Emergent features, attention, and object perception. *Journal of Experimental Psychology: Human Perception and Performance* **1984**, 10(1), 12.
54. Treisman, A. M., Gelade, G. A feature-integration theory of attention. *Cognitive Psychology* **1980**, 12(1), 97–136.
55. Treisman, A. Visual attention and the perception of features and objects. *Canadian Psychology/Psychologie canadienne* **1994**, 35(1), 107.
56. Itti, L., Koch, C. Computational modelling of visual attention. *Nature Reviews Neuroscience* **2001**, 2(3), 194–203.

57. Borji, A., Itti, L. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2012**, 35(1), 185–207.
58. Itti, L., Koch, C., Braun, J. Revisiting spatial vision: Toward a unifying
59. Itti, L. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing* **2004**, 13(10), 1304–1318.
60. Itti, L. Quantitative modelling of perceptual salience at human eye position. *Visual Cognition* **2006**, 14(4-8), 959–984.
61. Carmi, R., Itti, L. The role of memory in guiding attention during natural vision. *Journal of Vision* **2006**, 6(9), 4–4.
62. Itti, L., Gold, C., Koch, C. Visual attention and target detection in cluttered natural scenes. *Optical Engineering* **2001**, 40(9), 1784–1793.
63. Itti, L., Koch, C. Comparison of feature combination strategies for saliency-based visual attention systems. In *Human Vision and Electronic Imaging IV* (1999), SPIE, 3644, 473–482.
64. Zhang, Y., Wu, X., Zheng, C., et al. Effects of vergence eye movement planning on size perception and early visual processing. *Journal of Cognitive Neuroscience* **2024**, 1–14.
65. Yuxuan, C., Zhang, D., Liang, B., et al. Visual creativity imagery modulates local spontaneous activity amplitude of resting-state brain.
66. Zhang, W., He, X., Liu, S., et al. Neural correlates of appreciating natural landscape and landscape garden: Evidence from an fMRI study. *Brain and Behavior* **2019**, 9(7), e01335.
67. Wen, X., Zhang, D., Liang, B., et al. Reconfiguration of the brain functional network associated with visual task demands. *PLoS One* **2015**, 10(7), e0132518.
68. Zhou, L. F., Wang, K., He, L., et al. Twofold advantages of face processing with or without visual awareness. *Journal of Experimental Psychology: Human Perception and Performance* **2021**, 47(6), 784.
69. Zhang, Y., Wu, X., Zheng, C., et al. Effects of vergence eye movement planning on size perception and early visual processing. *Journal of Cognitive Neuroscience* **2024**, 1–14.
70. Zhu, H., Cai, T., Xu, J., et al. Neural correlates of stereoscopic depth perception: a fNIRS study. In *2016 Progress in Electromagnetic Research Symposium (PIERS)* (2016), IEEE, 4442–4446.
71. Li, J., Chen, S., Wang, S., Lei, M., Dai, X., Liang, C., Xu, K., Lin, S., Li, Y., Fan, Y., et al. "An optical biomimetic eyes with interested object imaging." *arXiv*, 2021, arXiv:2108.04236.
72. Li, J., Lei, M., Dai, X., Wang, S., Zhong, T., Liang, C., Wang, C., Xie, P., Wang, R. "Compressed Sensing Based Object Imaging System and Imaging Method Therefor." U.S. Patent 2021/0144278 A1, 13 May 2021.
73. Ronneberger, O., Fischer, P., Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* (2015), Springer International Publishing, 234–241.
74. Rosenblatt, Frank. "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological Review* 65, no. 6 (1958): 386–408.
75. Liu, Ziming, et al. "KAN: Kolmogorov-Arnold Networks." *ArXiv abs/2404.19756* (2024): n. pag.
76. Zhang, K., Zuo, W. M., and Zhang, L. "Deep plug-and-play super-resolution for arbitrary blur kernels." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C.: IEEE Computer Society, 2019, pp. 1671–1681.
77. Horé, A., and D. Ziou. "Image Quality Metrics: PSNR vs. SSIM." In *2010 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010, pp. 2366–2369. doi: 10.1109/ICPR.2010.579.
78. Bauer, Eric, and Ron Kohavi. "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants." *Machine Learning* 36 (1999): 105–139.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.