Computational Evidence for the Functional Necessity of Marr's 2.5D Sketch and Gestalt Principles in 3D Scene Perception

Yixu Huang[1] **& Rui Zhong**[†1]
[1]School of Computer Science, Central China Normal University, Wuhan, China
[†]Corresponding author: zhongr@ccnu.edu.cn

Abstract

The human visual system robustly reconstructs 3D scenes from fleeting glances and sparse 2D inputs, effectively solving an ill-posed problem. In contrast, machine vision systems like 3D Gaussian Splatting often fail under such conditions, suffering from severe overfitting. We hypothesize this gap stems from the absence of biologically plausible inductive biases: specifically, Marr's 2.5D sketch and Gestalt organizational principles. To test this, we operationalized these theories within a computational framework, incorporating surface orientation cues (Marr) and the Gestalt principle of Prägnanz to enforce structural simplicity and suppress redundancy. Experiments demonstrate that these cognitive constraints are computationally necessary to resolve geometric ambiguities in sparse-view environments. Crucially, our biologically constrained model yields representations judged by humans as significantly more realistic and structurally coherent than unconstrained baselines. These findings provide compelling computational evidence supporting the functional necessity of intermediate geometric representations and global organization principles in facilitating robust human 3D perception.

Keywords: Marr's 2.5D sketch; Gestalt principles; 3D Reconstruction; Visual Cognition; Geometric Perception

Introduction

Visual perception presents a fundamental puzzle in cognitive science known as the inverse optics problem (Kawato et al., 1993). The human visual system receives only sparse and two-dimensional retinal inputs yet robustly reconstructs comprehensive three-dimensional representations of scenes involving geometric configurations and volumetric data. This proficiency implies that human vision solves a mathematically ill-posed problem by relying on strong biological constraints and priors. In stark contrast, contemporary machine vision systems exhibit severe performance degradation when restricted to information-limited conditions such as sparse viewpoints. This divergence suggests that the absence of biologically plausible inductive biases limits the cognitive robustness of current artificial systems. Consequently, computational modeling that embeds specific visual cognition mechanisms into machine architectures offers a dual opportunity to advance artificial perception and to provide computational evidence for the functional necessity of these biological theories.

To bridge the gap between two-dimensional sensation and three-dimensional perception, David Marr's computational theory (Kitcher, 1988; Marr, 2010) posits the existence of a 2.5D sketch (Wu et al., 2017). This intermediate and observer-centric representation encodes surface depth and orientation before full object recognition occurs, acting as a critical structural scaffold to resolve geometric ambiguities. Complementing this, Gestalt theory (Arnheim, 1949; Mungan, 2023) addresses the organization of visual information by asserting that global configurations emerge through grouping principles such as continuity and the Law of Prägnanz (Van Geert and Wagemans, 2024) or simplicity. While Marr's theory emphasizes the hierarchical construction of geometry, Gestalt principles highlight the suppression of redundancy and the coherence of the whole. We hypothesize that these two cognitive mechanisms are not merely descriptive but are computationally necessary inductive biases for robust 3D scene understanding. Current machine vision approaches often lack these constraints and consequently exhibit overfitting behaviors under sparse input conditions by generating spurious information in occluded regions.

Figure 1: The comparison between our method and the 3DGS method demonstrates that our approach, by simulating human visual cognition, achieves superior viewpoint synthesis performance through enhanced perception of 3D scenes.

To test this hypothesis, we utilize 3D Gaussian Splatting (3DGS, Kerbl et al., 2023) as a computational proxy for visual reconstruction. While 3DGS achieves state-of-the-art rendering by projecting discrete 3D points, it suffers from a lack of cognitive constraints. Under sparse-view conditions, the ill-posed nature of the task causes the unconstrained model to generate hallucinations in the form of excessive Gaussian points and physically implausible artifacts. This mirrors perceptual errors that occur in the absence of valid priors. Existing engineering solutions (Chung et al., 2024; Li et al., 2024; Zhu et al., 2024) attempt to address this with monocular depth estimation (Yang, Kang, Huang, Xu, et al., 2024; Yang, Kang, Huang, Zhao, et al., 2024) but often fail to fix the underlying structural incoherence and redundancy. This dependency

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

leads to artifacts that deviate from human perceptual expectations. Therefore, integrating cognitive theories such as Marr's 2.5D sketch and Gestalt principles offers a promising pathway to regularize machine perception in a human-like manner.

We propose a biologically constrained computational framework that operationalizes Marr's and Gestalt theories to regularize 3D scene reconstruction. To simulate the human capacity for perception via fleeting glances, we restrict inputs to sparse images. First, inspired by Marr's 2.5D sketch, we introduce a global-gradient-local collaborative depth regularization strategy. This mechanism enforces a coherent intermediate representation to guide the geometric learning process and mimics the role of the 2.5D sketch in bridging raw input and 3D structure. Second, drawing on Gestalt principles of simplicity and organization, we implement a Gaussian control regularization approach. By imposing penalties on opacity and structural complexity, we mathematically model the visual system's tendency to suppress redundant information and noise. This approach continuously enforces the principle of simplicity during optimization rather than relying on retrospective pruning strategies.

As demonstrated in Figure 1, our bio-inspired approach yields representations that are structurally coherent and free of the artifacts plaguing unconstrained models. In summary, our contributions are as follows:

- We propose a multi-dimensional depth regularization approach by simulating Marr's 2.5D sketch theory. This method effectively enhances 3D perception capabilities and demonstrates the computational utility of biologically inspired intermediate representations.
- We develop a multi-attribute Gaussian control approach based on Gestalt theory. This technique mimics the human visual system's ability to suppress redundant information and successfully eliminates superfluous artifacts during the optimization process.
- Experimental results demonstrate that our human-cognition-inspired 3DGS method achieves state-of-the-art performance. These findings provide computational evidence supporting the functional necessity of both Marr's 2.5D sketch and Gestalt organization principles as essential inductive biases for resolving the ambiguity of 3D perception.

Related Work

Recent advances in 3D cognition research have increasingly focused on the synergy between biological plausibility and computational efficiency. Here, we review the theoretical foundations of human visual processing and their specific relevance to modern neural rendering architectures.

Marr's Computational Framework and Mid-Level Vision. David Marr's seminal framework posits that visual processing is fundamentally hierarchical and transforms 2D retinal intensities into 3D representations through distinct algorithmic stages. Central to this architecture is the *2.5D sketch*. This intermediate and viewer-centric representation explicitly encodes surface orientation, depth maps, and discontinuities before high-level object recognition occurs. In cognitive science, this stage is considered a functional necessity for bridging the epistemic gap between low-level primal sketches and full volumetric understanding. While modern computer vision has often bypassed this structured stage in favor of end-to-end learning, we argue that the 2.5D sketch provides essential inductive biases. It effectively constrains the solution space for geometry recovery, particularly in tasks where surface reconstruction from monocular cues presents a classic inverse optics problem that is mathematically ill-posed.

Gestalt Principles as Computational Priors. Complementing Marr's structural approach, Gestalt theory offers a mechanism for perceptual organization and global coherence. It asserts that the visual system overcomes sensory sparsity by grouping local stimuli into unified global entities based on laws such as continuity, similarity, and *Prägnanz* or simplicity. From a computational perspective, these principles function as powerful regularization terms or priors that suppress information redundancy and enforce cognitive economy. They explain how biological vision infers complete structures from occluded or sparse inputs by prioritizing global structural integrity over local irregularities. In our framework, we operationalize these psychological principles to mitigate overfitting in machine learning models. We interpret the good form of Gestalt theory as a constraint against the generation of spurious artifacts in scene reconstruction, thereby aligning algorithmic optimization with human perceptual tendencies.

3D Gaussian Splatting as a Cognitive Testbed. In the domain of neural rendering, 3D Gaussian Splatting (3DGS) represents a paradigm shift from implicit neural fields (Mildenhall et al., 2020) to explicit and point-based representations (Qi et al., 2017; Shi et al., 2019). Kerbl et al. model scenes as a cloud of anisotropic 3D Gaussians which are optimized via differentiable rasterization to capture complex visual stimuli. While 3DGS achieves high-fidelity rendering, it inherently acts as a tabula rasa that lacks the structural constraints of the human visual system. Under sparse-view conditions, unconstrained 3DGS suffers from severe overfitting which manifests as visual hallucinations or geometric anomalies due to the absence of valid geometric priors. Consequently, 3DGS serves as an ideal computational testbed to validate the necessity of cognitive constraints. By injecting Marr's geometric cues and Gestalt-based regularization into this architecture, we can empirically test their functional role in stabilizing 3D perception under uncertainty.

Computational Modeling Framework

Preliminary: 3D Gaussian Splatting as a Baseline Representation

To computationally model the reconstruction of 3D scenes, we utilize 3D Gaussian Splatting (3DGS) as our baseline representation. This choice represents a paradigm shift from implicit neural representations to an explicit and point-based

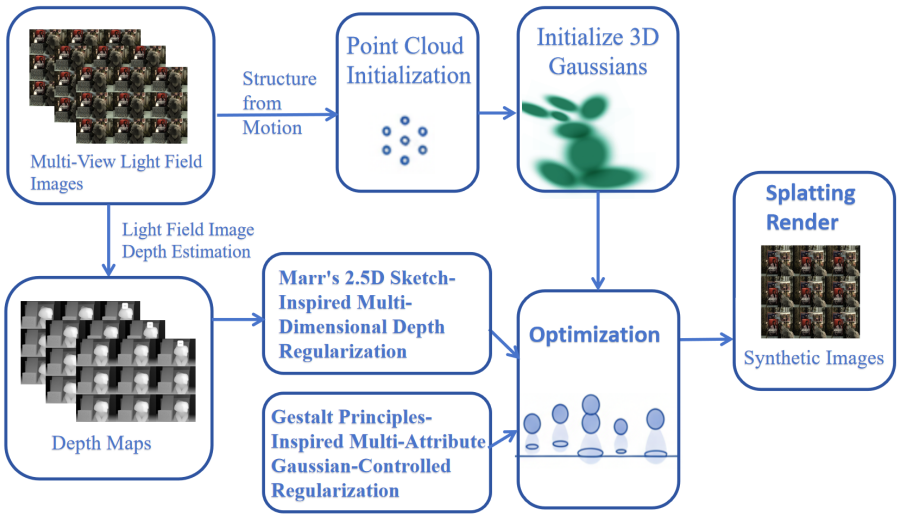

Figure 2: Overview of the proposed method. Our method begins by taking multi-view light field images as input and performing depth estimation. We utilize Structure from Motion (SfM) to generate an initial point cloud, which serves as the initialization for the 3D Gaussians. To guide the optimization of the 3DGS model, we incorporate two regularization terms: Marr's 2.5D sketch-inspired multi-dimensional depth regularization and Gestalt principles-inspired multi-attribute Gaussian-controlled regularization. Finally, high-quality novel views are synthesized via rendering.

scene representation. Unlike voxel grids, 3DGS models a scene as a collection of 3D Gaussians which can be conceptualized as a cloud of soft volumetric ellipsoids. This explicit structure combines the flexibility of point clouds with continuous differentiability and makes it uniquely suitable for capturing complex visual stimuli.

Mathematically, the scene is parameterized by a set of Gaussians where each is defined by a center $\boldsymbol{\mu} \in \mathbb{R}^3$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$ describing its shape and orientation. The influence of a Gaussian at a spatial point $\boldsymbol{x}$ follows the standard multivariate equation, as shown in (1):

$$G(\boldsymbol{x}) = \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{1}$$

To maintain physical validity during optimization, $\boldsymbol{\Sigma}$ is decomposed into rotation $\boldsymbol{R}$ and scaling $\boldsymbol{S}$ matrices such that $\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{R}^\top$. In addition to geometry, each Gaussian carries an opacity scalar α and Spherical Harmonics (SH) coefficients which encode view-dependent color effects.

The rendering process acts akin to projecting 3D brushstrokes onto a 2D canvas. To render an image, the 3D Gaussians are projected into 2D screen space. The 2D covariance $\boldsymbol{\Sigma}'$ is approximated using the viewing transformation $\boldsymbol{W}$ and the Jacobian of the affine projection $\boldsymbol{J}$, as shown in (2):

$$\boldsymbol{\Sigma}' = \boldsymbol{J}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^\top\boldsymbol{J}^\top. \tag{2}$$

Finally, pixel colors are computed using α-blending. The projected Gaussians covering a pixel are sorted by depth and their contributions are accumulated to form the final color C, as shown in (3):

$$C = \sum_{i\in\mathcal{N}} c_i\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{3}$$

where c_i is the color and α_i is the effective opacity of the i-th Gaussian. This differentiable pipeline allows the model to learn parameters solely by minimizing reconstruction error. However, without cognitive constraints, this optimization is prone to overfitting in sparse data regimes which mirrors the inverse optics problem in human vision.

Operationalizing Marr's 2.5D Sketch: Multi-Dimensional Depth Regularization

Marr's theory of vision posits that the human visual system constructs a 2.5D sketch to bridge the gap between 2D retinal inputs and 3D models. This intermediate representation explicitly encodes surface depth and orientation. To operationalize this theory within our computational framework, we introduce a multi-dimensional depth regularization strategy. This approach enforces structural constraints on the learning process and simulates the biological necessity of intermediate geometric representations.

Light field images, which capture both spatial and angular information, enable more accurate simulation of human visual perception and encode abundant geometric data. Therefore, we adopt light field images as input modalities in this study. As illustrated in Figure 2, our method leverages the rich information embedded in light field images. Specifically, we extract high-fidelity depth maps from sub-aperture images (SAIs) to serve as the ground truth for the 2.5D sketch. We propose three regularization components to align the learned geometry with this biological prior:

1. Depth Consistency (Global Constraint). To ensure the global geometric distribution aligns with the primal sketch of the scene, we employ a depth consistency loss. Follow-

ing LF-GS (Huang et al., 2025), we utilize Video Depth Anything (Chen et al., 2025) to extract depth maps by treating sub-aperture images as a pseudo-video sequence. This method fully exploits the rich geometric information inherent in light field images and yields more accurate depth maps than monocular depth estimation. Following standard protocols (Kerbl et al., 2024; Ranftl et al., 2022), we align the scale of these depths using COLMAP (Schonberger and Frahm, 2016)-derived sparse points. The constraint is formalized as the mean absolute error between the estimated inverse depth $\mathbf{D}^*$ and the rendered inverse depth $\hat{\mathbf{D}}$, as shown in (4):

$$\mathcal{L}_D = \frac{1}{N}\sum_{i=1}^{N} |\hat{\mathbf{D}}_i - \mathbf{D}_i^*|, \tag{4}$$

where i denotes the index of inverse depth values and N denotes the total number of samples.

2. Surface Orientation (Gradient Constraint). Marr emphasizes that the 2.5D sketch must encode surface orientation and discontinuities. To model this, we compute a depth gradient loss. This term constrains the gradient variation of the depth map to align with physical scene structures and prevents unreasonable abrupt changes in non-edge regions. It ensures that the system perceives continuous surfaces rather than disjointed points, expressed as (5):

$$\mathcal{L}_{grad} = \frac{1}{N}\sum_{i=1}^{N} \left|\nabla\hat{\mathbf{D}}_i - \nabla\mathbf{D}_i^*\right|. \tag{5}$$

3. Local Coherence (Statistical Constraint). To enforce stability within local receptive fields for depth maps, we introduce a local regularization strategy. This constraint regulates the statistical properties including mean and variance within a local window and thereby suppresses sensory noise and excessive dispersion. For each spatial location, we compute statistics within a 3×3 window Ω, expressed as (6), (7), and (8):

$$\mu_D(x,y) = \frac{1}{9}\sum_{(m,n)\in\Omega} \mathbf{D}(m,n), \tag{6}$$

$$\sigma^2_{\mathbf{D}}(x,y) = \frac{1}{9}\sum_{(m,n)\in\Omega} \left(\mathbf{D}(m,n) - \mu_{\mathbf{D}}(x,y)\right)^2, \tag{7}$$

$$\mathcal{L}_{local} = \frac{1}{N}\sum_{x,y}\left(\left|\mu_{\hat{\mathbf{D}}} - \mu_{\mathbf{D}^*}\right| + \left|\sigma^2_{\hat{\mathbf{D}}} - \sigma^2_{\mathbf{D}^*}\right|\right). \tag{8}$$

Operationalizing Gestalt Principles: Gaussian Control Regularization

Gestalt psychology suggests that human perception is governed by principles of grouping and the *Law of Prägnanz* which prioritizes simplicity, regularity, and order. In contrast, unconstrained machine optimization is susceptible to noise-induced overfitting and often results in complex but incoherent structures. We propose a multi-attribute Gaussian control regularization to operationalize these principles of perceptual organization.

1. Simplicity via Opacity Regulation (Law of Prägnanz). To enforce the principle of simplicity, we penalize the accumulation of ambiguous semi-transparent structures. By imposing a penalty on opacity, we encourage the model to resolve perceptual uncertainty by rendering insignificant Gaussians as completely transparent. This reduces representational redundancy, as shown in (9):

$$\mathcal{L}_o = \frac{1}{n}\sum_{i=1}^{n} \alpha_i, \tag{9}$$

where α_i is the opacity and n is the total count.

2. Compactness via Scale Constraint. To preserve geometric plausibility and validity, we impose constraints on the spatial extent of scene elements. This prevents the unbounded growth of Gaussians that would otherwise cover irrelevant regions and cause blurring, as shown in (10):

$$\mathcal{L}_s = \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{s_i}\|_2, \tag{10}$$

where $\boldsymbol{s_i}$ is the scaling factor.

3. Cognitive Economy via Count Control. Finally, we directly enforce a principle of cognitive economy by controlling the total number of primitives used to represent the scene. This prevents the excessive growth of representational units induced by overfitting, as shown in (11):

$$\mathcal{L}_{count} = \log(\boldsymbol{N}_{\text{Gauss}}), \tag{11}$$

where $\boldsymbol{N}_{\text{Gauss}}$ denotes the number of Gaussians.

Optimization Objective

The total loss function integrates these data terms and cognitive priors as formulated in (12), (13), and (14):

$$\mathcal{L}_M = \lambda_1 \cdot \mathcal{L}_D + \lambda_2 \cdot \mathcal{L}_{grad} + \lambda_3 \cdot \mathcal{L}_{local}, \tag{12}$$

$$\mathcal{L}_N = \lambda_4 \cdot \mathcal{L}_o + \lambda_5 \cdot \mathcal{L}_s + \lambda_6 \cdot \mathcal{L}_{count} \tag{13}$$

$$\mathcal{L} = \lambda_7 \cdot \mathcal{L}_1 + \lambda_8 \cdot \mathcal{L}_{SSIM} + \mathcal{L}_M + \mathcal{L}_N, \tag{14}$$

where λ parameters represent the weighting of these distinct constraints. $\mathcal{L}_1$ and $\mathcal{L}_{SSIM}$ are the standard reconstruction losses used in 3DGS.

Experimental Validation

Simulating Visual Input and Environmental Constraints

To simulate the epistemic challenge faced by the human visual system, we model the scene reconstruction task under conditions of information sparsity. We utilize light field images to approximate the rich retinal input available to biological vision while artificially restricting the number of observations to replicate the fleeting glance scenario. Specifically, we employ the LF-GS (Huang et al., 2025) dataset which captures scenes using a 3×3 array of Lytro Illum cameras. Notably, this is currently the only dataset that is optimally tailored to the specific requirements of our experimental setup. This hardware setup ensures simultaneous capture and effectively

Table 1: Quantitative comparisons of different methods.

Method	PSNR↑	SSIM↑	LPIPS↓
3DGS (Kerbl et al., 2023)	21.13	0.695	0.248
Mip-Splatting (Yu et al., 2024)	21.34	0.696	0.242
Pixel-GS (Z. Zhang et al., 2024)	20.92	0.688	0.257
MS-GS (Yan et al., 2024)	21.16	0.697	0.245
DR-GS (Chung et al., 2024)	21.43	<u>0.712</u>	0.243
DN-GS (Li et al., 2024)	21.49	0.701	<u>0.235</u>
Scaffold-GS (Lu et al., 2024)	<u>21.51</u>	0.702	0.248
DropGaussian (Park et al., 2025)	21.48	0.702	0.238
Ours	**22.11**	**0.732**	**0.218**

The best metric is in **bold**, second-best in <u>underlined</u>.

Table 2: Ablation Study of Our Method.

Setting	PSNR↑	SSIM↑	LPIPS↓
3DGS	21.13	0.695	0.248
w/o DR	21.79	0.722	0.228
w/o DGR	22.00	0.728	0.222
w/o LDR	21.98	0.728	0.223
w/o GOR	21.75	0.723	0.221
w/o GSR	22.03	0.729	0.221
w/o GCR	21.97	0.720	0.224
Ours	**22.11**	**0.732**	**0.218**

The best metric is shown in **bold**.

eliminates temporal artifacts such as motion blur or lighting shifts that are untypical of human saccades. Consequently, this dataset provides a robust testbed for analyzing how visual systems resolve geometric ambiguities from sparse binocular or multi-view cues.

Evaluation Metrics as Perceptual Proxies

We employ a hierarchy of objective metrics to evaluate the model's capability in perceiving 3D scenes and serve as computational proxies for human visual assessment. Peak Signal-to-Noise Ratio (PSNR) is utilized to quantify low-level reconstruction fidelity by calculating the ratio between the maximum signal power and corrupting noise. While standard in engineering, it focuses primarily on pixel-level accuracy. To capture higher-order perceptual qualities, we employ the Structural Similarity Index Measure (SSIM, Wang et al., 2004). This metric models the human visual system's sensitivity to structural information including luminance and contrast rather than absolute pixel errors. Furthermore, we utilize Learned Perceptual Image Patch Similarity (LPIPS, R. Zhang et al., 2018) which measures perceptual distance using deep features extracted from neural networks. LPIPS aligns more closely with human perceptual judgment regarding complex distortions and texture plausibility.

Implementation Details

To ensure a fair comparison with unconstrained baseline models, we retain learning rates and optimization parameters identical to the original 3DGS framework. The model undergoes 30,000 training iterations. Our loss coefficients are configured as follows: $\lambda_4 = 0.05$, $\lambda_5 = 0.008$, $\lambda_6 = 0.001$, $\lambda_7 = 0.8$, and $\lambda_8 = 0.2$. During the optimization process, depth-related losses including depth, depth gradient, and local depth utilize an exponentially decaying weighting strategy, where the respective weights λ_1, λ_2 and λ_3 start at 1 and decay to 0.01 for per-chunk optimization. This mimics coarse-to-fine processing: strong structural priors guide early learning, relaxation enables fine-grained detail emergence. Dataset partitioning mirrors standard protocols established in 3DGS.

We compare our biologically inspired framework with state-of-the-art unconstrained methods including 3DGS (Kerbl et al., 2023), Mip-Splatting (Yu et al., 2024), Pixel-GS (Z. Zhang et al., 2024), MS-GS (Yan et al., 2024), DR-GS (Chung et al., 2024), DN-GS (Li et al., 2024), Scaffold-GS (Lu et al., 2024) and DropGaussian (Park et al., 2025). The average performance is illustrated in Table 1. Our method achieves significant improvements across PSNR, SSIM, and LPIPS metrics. Specifically, the PSNR is 0.60dB higher than Scaffold-GS (Lu et al., 2024). This performance gap highlights the efficacy of integrating geometric priors derived from light field data to address the ill-posed nature of sparse view synthesis. It suggests that biological constraints are not merely optional optimizations but are critical for robust performance in data-scarce environments.

Ablation Study: Testing the Necessity of Cognitive Modules

To empirically validate the functional necessity of the proposed cognitive mechanisms, we conducted a systematic ablation study. This process effectively lesions specific components of the model to observe the resulting deficits in 3D perception.

Functional Necessity of Marr's 2.5D Sketch. To test the hypothesis that an intermediate geometric representation is essential for robust vision, we systematically removed the components of our multi-dimensional depth regularization: depth regularization, depth gradient regularization, and local depth regularization (denoted as w/o DR, w/o DGR, and w/o LDR respectively). As evidenced in Table 2, the removal of any of these terms results in significant performance degradation. This decline supports the theoretical stance that raw retinal input alone is insufficient for 3D reconstruction without the structural scaffolding provided by the 2.5D sketch. Our model successfully incorporates these geometric priors to resolve ambiguities that unconstrained models cannot handled. These findings provide computational evidence that the 2.5D sketch serves as a necessary inductive bias for both human and machine vision systems.

Functional Necessity of Gestalt Organization. To val-

idate the role of perceptual organization principles, we performed an ablation by removing the Gaussian opacity, scaling, and count regularizations (denoted as w/o GOR, w/o GSR, and w/o GCR respectively). As shown in Table 2, the exclusion of these constraints leads to a marked decline in performance. This result indicates that without the explicit enforcement of the *Law of Prägnanz* or simplicity, the optimization process degenerates into overfitting and produces redundant or noisy representations. By enforcing constraints on opacity and compactness, our model mimics the cognitive capacity to suppress irrelevant information and perceive good figures. These findings corroborate the hypothesis that robust 3D scene perception fundamentally relies on Gestalt-like mechanisms to filter sensory noise and construct coherent scene representations.

Validation via Human Perception

While objective metrics provide quantitative benchmarks, the ultimate test of a cognitive model is its alignment with human perception. We conducted a blind user study to evaluate the subjective quality of the synthesized views. We invited 40 participants with diverse backgrounds to perform a two-alternative forced choice (2AFC) task. Participants were presented with randomized pairs of images rendered by our method and state-of-the-art baselines. They were instructed to select the result that appeared more realistic, structurally coherent, and visually comfortable (defined as having fewer artifacts). Our method was preferred in over 95 percent of the trials. Participants consistently reported that our results exhibited significantly fewer visual floaters and sharper geometric details compared to baseline methods particularly in sparse-view scenarios. This overwhelming preference supports our core hypothesis: incorporating Marr's depth cues and Gestalt-based redundancy suppression effectively eliminates the visual ambiguities that human observers find unnatural and results in 3D representations that are perceptually congruent with human vision.

Psychophysical Validation of Spatial Structural Integrity

To empirically validate the functional necessity of the proposed cognitive constraints beyond subjective visual preference, we conducted a psychophysical experiment using a relative depth discrimination task designed to probe the structural integrity and cognitive efficiency of the reconstructed scenes. Twenty participants were presented with stimuli from ground truth, an unconstrained baseline, and our biologically constrained method, and were asked to identify the spatially closer of two probe points positioned in geometrically ambiguous regions. This experimental design directly operationalizes Marr's theory by testing the fidelity of the recovered 2.5D sketch while simultaneously assessing Gestalt efficiency via processing speed. As shown in Figure 3, the results revealed that the unconstrained baseline yielded significantly lower accuracy and slower reaction times compared to ground truth,

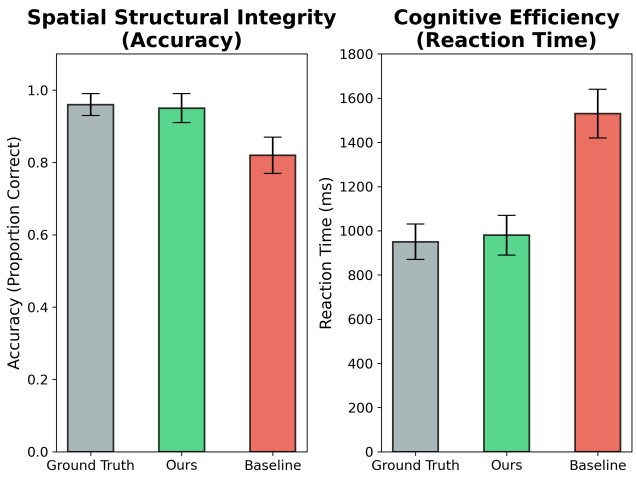

Figure 3: The results verified by psychophysics.

indicating that the absence of inductive biases leads to geometric hallucinations that disrupt spatial reasoning and increase cognitive load. In contrast, our method achieved accuracy and reaction times statistically indistinguishable from ground truth. These findings provide compelling computational evidence that incorporating Marr's surface orientation cues and Gestalt principles for redundancy suppression is functionally necessary to restore the spatial affordances of a scene, enabling human-like robustness in perceiving 3D structure from sparse inputs.

Conclusions and Discussions

This study presents a computational framework that operationalizes classical theories of visual cognition within the context of modern neural rendering. By integrating Marr's 2.5D sketch as a structural scaffold and employing Gestalt-inspired mechanisms for redundancy suppression, we provide computational evidence that biologically plausible constraints are essential for robust perception in sparse information environments. Our findings extend beyond technical metrics and offer support for the functional necessity of intermediate representations in the human visual hierarchy. Specifically, we demonstrate that the 2.5D sketch serves as a critical inductive bias which enables the visual system to resolve the inherent ambiguities of the inverse optics problem by bridging low-level sensory features and high-level volumetric understanding. This supports the constructivist view that human depth perception relies not merely on the passive integration of cues but on the active and constraints-driven construction of surface geometry. Furthermore, the efficacy of our multi-attribute regularization strategy suggests that the principle of cognitive economy or Prägnanz is mathematically advantageous for learning efficient and generalizable scene representations. Ultimately, this bio-inspired paradigm suggests that artificial systems can achieve human-like robustness by mimicking the architectural constraints of biological vision.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62002130 and 62377026, in part by the Fundamental Research Funds for Central China Normal University (Nos. CCNU25ai016 and CCNU26ai), and in part by the CCNU Digital Intelligence–Empowered Educational Innovation and Teaching Reform Project (No. CCNU25JG14).

References

Arnheim, R. (1949). The gestalt theory of expression. *Psychological Review*, *56*(3), 156.

Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., & Kang, B. (2025). Video depth anything: Consistent depth estimation for super-long videos. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22831–22840.

Chung, J., Oh, J., & Lee, K. M. (2024). Depth-regularized optimization for 3d gaussian splatting in few-shot images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 811–820.

Huang, Y., Zhong, R., Rogge, S., & Munteanu, A. (2025). Lf-gs: 3d gaussian splatting for view synthesis of multi-view light field images. *IEEE Signal Processing Letters*, *32*, 3555–3559.

Kawato, M., Hayakawa, H., & Inui, T. (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network: Computation in Neural Systems*, *4*(4), 415.

Kerbl, B., Kopanas, G., Leimkühler, T., & Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, *42*(4), 1–14.

Kerbl, B., Meuleman, A., Kopanas, G., Wimmer, M., Lanvin, A., & Drettakis, G. (2024). A hierarchical 3d gaussian representation for real-time rendering of very large datasets. *ACM Transactions on Graphics*, *43*(4), 1–15.

Kitcher, P. (1988). Marr's computational theory of vision. *Philosophy of Science*, *55*(1), 1–24.

Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., & Gu, L. (2024). Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20775–20785.

Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., & Dai, B. (2024). Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20654–20664.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). Nerf: Representing scenes as neural radiance fields for view synthesis. *European Conference on Computer Vision*, 405–421.

Mungan, E. (2023). Gestalt theory: A revolution put on pause? prospects for a paradigm shift in the psychological sciences. *New Ideas in Psychology*, *71*, 101036.

Park, H., Ryu, G., & Kim, W. (2025). Dropgaussian: Structural regularization for sparse-view gaussian splatting. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21600–21609.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 652–660.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2022). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(3), 1623–1637.

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Shi, S., Wang, X., & Li, H. (2019). Pointrcnn: 3d object proposal generation and detection from point cloud. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–779.

Van Geert, E., & Wagemans, J. (2024). Prägnanz in visual perception. *Psychonomic Bulletin & Review*, *31*(2), 541–567.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wu, J., Wang, Y., Xue, T., Sun, X., Freeman, B., & Tenenbaum, J. (2017). Marrnet: 3d shape reconstruction via 2.5 d sketches. *Advances in Neural Information Processing Systems*, *30*.

Yan, Z., Low, W. F., Chen, Y., & Lee, G. H. (2024). Multi-scale 3d gaussian splatting for anti-aliased rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20923–20931.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything v2. *Advances in Neural Information Processing Systems*, *37*, 21875–21911.

Yu, Z., Chen, A., Huang, B., Sattler, T., & Geiger, A. (2024). Mip-splatting: Alias-free 3d gaussian splatting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19447–19456.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhang, Z., Hu, W., Lao, Y., He, T., & Zhao, H. (2024). Pixel-gs: Density control with pixel-aware gradient for 3d gaussian splatting. *European Conference on Computer Vision*, 326–342.

Zhu, Z., Fan, Z., Jiang, Y., & Wang, Z. (2024). Fsgs: Real-time few-shot view synthesis using gaussian splatting. *European Conference on Computer Vision*, 145–163.