On the Origin of 3D Perception in Visual World Models

Wanhee Lee*, Klemen Kotar*, Jared Watrous*, Rahul Mysore Venkatesh*,
Honglin Chen, Khai Loong Aw, Khaled Jedoui, Daniel LK Yamins
Stanford University

Abstract

3D perception is essential for both biological and artificial vision, supporting navigation, object interaction, and scene understanding. However, learning 3D structure in a self-supervised manner is challenging because fully structured geometric methods impose rigid constraints that limit adaptability to natural videos, while unstructured, data-only approaches lack geometric consistency and struggle with controllability. We investigate a balanced approach that starts with minimal priors and progressively builds structured representations, using optical flow as an intermediate representation to infer depth and subsequently 3D shape, integrating flexibility with geometric consistency. Using this framework, we demonstrate that our model performs competitively on 3D perception tasks from a single image, achieving human-level depth estimation and supporting shape inference beyond visible surfaces. Beyond accuracy, we examine the model's biological relevance. Our results indicate that depth perception develops rapidly when learning from camera motion, similar to early visual learning in humans. These findings support a middle ground between structured and unstructured learning, providing a biologically plausible path for self-supervised 3D perception.

Keywords: 3D Perception; Computer Vision

Introduction

3D perception is fundamental to vision, enabling biological organisms to navigate, interact with objects, and understand their surroundings. It is also a key capability for artificial vision systems.

A central question in 3D vision is whether core components such as depth and shape perception require strong structural priors, for example geometric constraints and camera models, or whether they can emerge from data alone. Fully structured approaches rely on explicit geometric reasoning and 3D representations (Yu, Ye, Tancik, & Kanazawa, 2021; Tewari et al., 2023), but impose constraints that limit their applicability to natural video. Conversely, unstructured models (Z. Wang et al., 2024) aim to learn directly from raw observations, yet often struggle with geometric consistency and controllability.

We propose a hybrid framework that combines the flexibility of unstructured learning with the controllability of structured reasoning. An autoregressive video model is trained without geometric priors, using optical flow obtained from natural video as an intermediate cue. Structure is reintroduced at inference time by computing flow from estimated depth and camera motion, which enables controllable 3D inference.

*These authors contributed equally.

Our framework builds on Counterfactual World Modeling (CWM) (Bear et al., 2023), which showed that optical flow can be extracted from unstructured predictors. We extend this idea to demonstrate that higher-level 3D representations, specifically depth and shape, can also emerge.

We evaluate our model on two core tasks from a single image, depth estimation and shape completion through novel view synthesis, and analyze the developmental trajectory of depth learning. Our model achieves human-level depth estimation (Zuo et al., 2024), infers coherent 3D shape representations, and exhibits a learning trajectory that parallels human depth development.

Motion cues, particularly those arising from self-motion, play a critical role in the emergence of depth perception (Gibson & Walk, 1960). Our model provides a computational example of this process, showing that depth can emerge rapidly from motion-based learning signals and offering insights for both artificial and biological vision.

Method

Our model builds on the Probabilistic Structure Integrator (PSI), an autoregressive multimodal visual world model inspired by Counterfactual World Modeling (CWM) (Bear et al., 2023). PSI consists of an unstructured base predictor and a set of structured prompting programs that extract intermediate 3D representations.

The unstructured base predictor is an autoregressive transformer trained to predict the next RGB frame from the current frame. It is worth noting that optical flow can be extracted from this base model using structured prompting, following CWM. We train a flow predictor to estimate optical flow conditioned on camera pose changes, and a flow-to-RGB predictor, finetuned from the RGB predictor, that generates the next frame conditioned on flow and the input image. Conditioning RGB generation on flow provides controllable and geometrically consistent RGB generation.

We introduce structured prompting programs during inference to extract depth, perform novel view synthesis, and support 3D shape reconstruction. Depth is extracted by applying controlled in-plane camera motion, which induces optical flow whose magnitude is inversely related to scene depth. Novel view synthesis is performed by warping the depth map to obtain a target flow field, followed by RGB generation. For 3D shape extraction, we generate novel RGB views from a single input image and use these views as input to an external multi-view reconstruction method (S. Wang, Leroy, Cabon, Chidlovskii, & Revaud, 2024) to obtain full 3D geometry.

PSI was trained in a self-supervised manner on approximately one year of diverse, publicly available internet videos

and multi-view datasets with camera poses, including ScanNet++, CO3D, MVImgNet, and RealEstate10K. We used RAFT (Teed & Deng, 2020) to precompute optical flow, used to supervise the flow predictor and condition the flow-to-RGB predictor during training. The RGB model has 7 billion parameters and the flow predictor has 1 billion. Both models operate on tokenized RGB and flow representations. During inference, PSI models generate tokens autoregressively, conditioned on RGB, camera poses, or flow fields.

Results

We evaluate our model's capacity for 3D understanding through two core tasks, monocular depth estimation and shape completion via novel view synthesis, and analyze how depth perception emerges during training.

Monocular Depth Estimation

We first assess the model's ability to estimate depth from a single image, using the UniQA-3D benchmark (Zuo et al., 2024). This task involves determining which of two points in an image is farther away, providing a cognitively relevant measure of depth understanding. Figure 1 shows that our model achieves human-level accuracy on upright images and generalizes well to flipped images, demonstrating robustness in geometric reasoning. These results support the hypothesis that structured depth representations can emerge from a minimally structured base model guided by optical flow cues.

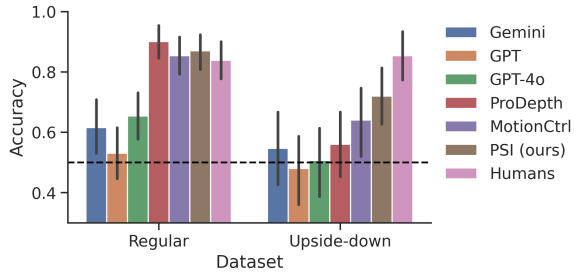

Figure 1: **Quantitative Evaluation of Depth Estimation on UniQA-3D.** Our model matches human-level performance on upright images and generalizes well to flipped images, indicating robust depth understanding.

3D Shape Understanding through Novel View Synthesis

To evaluate the model's ability to infer complete 3D structure, we use novel view synthesis as an intermediate step toward 3D shape completion. Given a single input image, we first extract depth, then synthesize novel views at new camera poses, and finally use these views to reconstruct 3D shape with an external multi-view method (S. Wang et al., 2024).

Figure 2 illustrates how our model's synthesized novel views reveal occluded scene regions and object parts, enabling more complete 3D reconstruction. This demonstrates that the model captures not only visible surfaces but also plausible amodal structure, a key feature of human 3D perception.

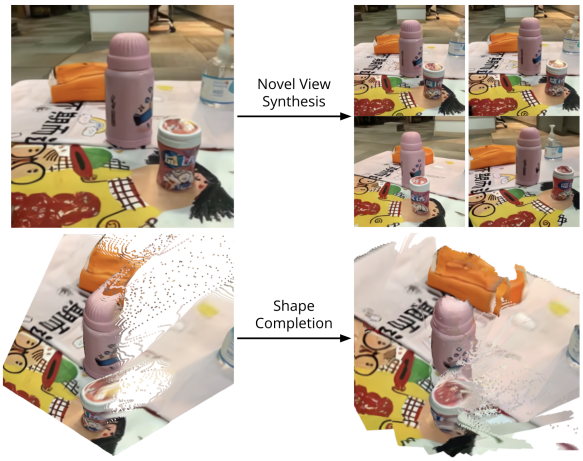

Figure 2: **3D Shape Completion through Novel View Synthesis.** Novel views synthesized from a single image reveal occluded regions and enable more complete shape reconstruction, supporting amodal 3D understanding.

Developmental Plausibility

Finally, we analyze the developmental trajectory of depth perception in our model. Motion cues are known to contribute to depth learning in biological systems (Gibson & Walk, 1960; Held & Hein, 1963). We measure how depth understanding emerges during training and find that the model acquires robust depth estimation early in training, after limited data exposure corresponding to about one month of awake time. This is consistent with the rapid development of depth perception in biological systems and highlights the effectiveness of optical flow as a learning signal for 3D structure.

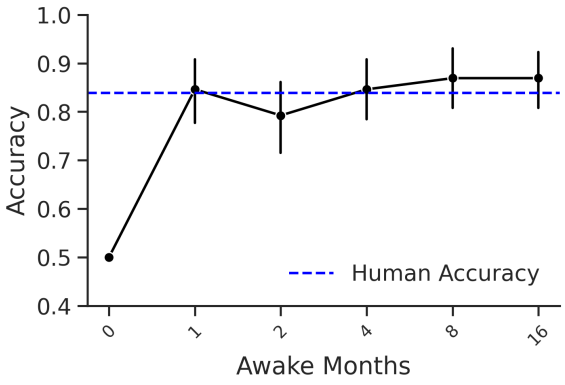

Figure 3: **Developmental trajectory of depth perception.** Depth understanding emerges early in training with limited data in a flow-based learning framework, consistent with the rapid development of depth perception in biological systems.

Conclusions

Our results show that 3D perception can emerge from a hybrid model combining unstructured learning with structured optical flow and depth reasoning. This work offers a step toward building models that acquire 3D scene understanding from natural video in a developmentally plausible way.

References

Bear, D. M., Feigelis, K., Chen, H., Lee, W., Venkatesh, R., Kotar, K., ... Yamins, D. L. (2023). Unifying (machine) vision via counterfactual world modeling. *arXiv preprint arXiv:2306.01828*.

Gibson, E. J., & Walk, R. D. (1960). The "visual cliff". *Scientific American*, *202*(4), 64–71.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. *J. Comp. and Physiol. Psychology*, *56*(5), 872–876.

Teed, Z., & Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. In *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part ii 16* (pp. 402–419).

Tewari, A., Yin, T., Cazenavette, G., Rezchikov, S., Tenenbaum, J. B., Durand, F., ... Sitzmann, V. (2023). Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *NeurIPS*.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., & Revaud, J. (2024). Dust3r: Geometric 3d vision made easy. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 20697–20709).

Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., & Shan, Y. (2024). Motionctrl: A unified and flexible motion controller for video generation. In *Siggraph.*

Yu, A., Ye, V., Tancik, M., & Kanazawa, A. (2021). pixelNeRF: Neural radiance fields from one or few images. In *Cvpr.*

Zuo, Y., Kayan, K., Wang, M., Jeon, K., Deng, J., & Griffiths, T. L. (2024). Towards foundation models for 3d vision: How close are we? *arXiv preprint arXiv:2410.10799*.