MonoPop3D: Single-Image Volumetric 3D Pop-up Generation

Yunyi Shi, Ruizhe(Rachel) Tan, Lynn Li

Abstract—Single-view 3D reconstruction often struggles with complex, non-rigid subjects such as furry pets, resulting in flat geometry and jagged boundary artifacts. We present *MonoPop3D*, a computational photography pipeline that transforms a single RGB image into a volumetric 3D pop-up animation without requiring multi-view constraints. Our hybrid approach integrates deep learning priors for semantic understanding with test-time mathematical optimization for visual refinement. Specifically, we employ interactive segmentation (SAM) with topological repair to resolve occlusion, followed by Guided Filter alpha matting to recover semi-transparent details like fur. To simulate realistic 3D rotation from a single view, we introduce a non-linear volumetric warping method driven by monocular depth estimation, which exaggerates internal curvature. Furthermore, we propose novel image-domain enhancements, for example normal-based relighting and depth-dependent bokeh, to reinforce the perception of 3D shape and focus. Qualitative and quantitative evaluations demonstrate that our method significantly reduces boundary artifacts and improves volumetric perception compared to standard planar parallax techniques.

Index Terms—Computational Photography, Single-View 3D, Alpha Matting, Generative Inpainting, Volumetric Parallax, Interactive Segmentation.

1 Introduction

Single-image 3D visualization technology has recently garnered significant attention across education, digital storytelling, virtual museums, and creative media. Traditional 2D images no longer satisfy users because they lack true geometric depth, leading to flat visual effects. While full 3D reconstruction techniques like Multi-View Stereo (MVS) and Neural Radial Field (NeRF) can restore precise geometric structures [1], they require multi-view photography, controlled environments, and massive computational resources. These limitations hinder their adoption by general users and real-time interactive scenarios. Consequently, lightweight solutions that simulate 3D perception effects from a single image are increasingly favored.

The existing single-image "2.5D" parallax and photo animation techniques suffer from several limitations. First, most systems assume a planar foreground model, generating a cardboard-cutout effect. Objects fail to convey internal depth variations and exhibit rigid sliding during motion. Second, foreground extraction often relies on binary masks. When handling fuzzy boundaries like hair, fur, or soft materials, edges appear jagged and flicker dynamically. Third, classical diffusion-based inpainting techniques often produce blurred textures, leading to inconsistent background structures. This becomes particularly noticeable during animation playback. These limitations diminish the realism of three-dimensional motion, thereby compromising the visual experience.

To address these limitations, we propose MonoPop3D. Our approach is a hybrid computational photography pipeline that generates a volumetric 3D pop-up animation from a single RGB image, without requiring multiple viewpoints. Our method does not perform full geometric reconstruction. It focuses on improving visual depth using pretrained deep models and test-time optimization. We combine pretrained deep models for interactive segmentation (SAM) [2], monocular depth estimation (MiDaS), and background inpainting (LaMa) [3] with classic image processing methods, including morphological topology repair, guided-filter alpha matting [4], and non-linear depth-based volumetric warping. This hybrid design provides reliable semantic understanding and precise control over object boundaries and depth deformation.

- *Yunyi Shi, Ruizhe(Rachel) Tan and Lynn Li are with University of Toronto.*

2 Related Work

2.1 Layered Depth Images

Shade et al. introduced the Layered Depth Images (LDI) to store multiple depth samples per pixel [5]. This technique synthesizes the view and handles occlusion. By separating foreground and background layers, LDI-based rendering enables limited parallax effects from a single reference view. This idea later became a core component in many image-based rendering systems. However, classical LDI methods rely heavily on accurate depth maps and clean object boundaries, which are difficult to obtain from a single RGB image. In addition, LDI pipelines assume piecewise-planar geometry, leading to rigid motion and failing to capture internal volumetric structure. As a result, rendered results often exhibit visible layer separation artifacts and lack realistic depth deformation.

MonoPop3D adopts the layered foreground/background representation of LDI for robustness and efficiency. However, instead of treating the foreground as a rigid plane, we introduce non-linear volumetric warping based on monocular depth. Our method enables internal depth variation and volumetric motion. Furthermore, we integrate alpha matting and deep inpainting to improve boundary

quality and background consistency beyond what classical LDI methods provide.

2.2 Depth-Based Single-Image 3D Photography

Shih et al. introduced a single-image 3D photography framework that generates parallax-based animations using context-aware layered depth inpainting [6]. Their method predicts monocular depth from a single RGB image, separates the scene into foreground and background layers, and applies depth-guided background inpainting to synthesize novel camera views. This approach enables convincing 3D motion from a single photograph and has become a foundational method for image-based 3D photo animation.

Despite its strong visual performance, these methods have several limitations. First, foreground extraction relies on binary segmentation masks, which introduce rigid boundaries and fail on fuzzy regions such as hair and fur. Second, they use early deep inpainting methods to complete the background. This approach may blur textures and lack global structural consistency when large regions are missing. Third, the depth-based warping uses linear or planar motion, which causes objects to move as rigid cutouts and limits the perception of internal volume.

MonoPop3D is most closely related to the method of Shih et al., but extends it in three critical directions. We replace binary masks with guided-filter alpha matting to recover soft boundaries, we adopt LaMa-based Fourier inpainting to improve large-scale background consistency, and we introduce non-linear depth-driven volumetric warping to produce internal parallax rather than planar motion. These extensions directly address the core failure modes of classical depth-based 3D photo animation.

2.3 Neural Rendering and NeRF-Based Methods

Neural Radiance Fields (NeRF) and related neural rendering methods learn continuous 3D scene representations from multiple images [1]. These methods synthesize new views by modeling scene geometry and appearance, achieving very high realism. Thus, they produce accurate geometry, lighting, and occlusion effects.

However, NeRF-based methods require dozens to hundreds of input views, long per-scene optimization time, and significant GPU memory. They are therefore unsuitable for single-image input, real-time interaction, or casual user applications. In addition, NeRF focuses on accurate 3D reconstruction rather than fast perceptual animation.

In contrast, we design MonoPop3D for single-image, zero-shot inference using pretrained models. While it does not attempt to reconstruct full geometry, it produces real-time volumetric depth cues suitable for lightweight pop-up animation and interactive applications.

3 Proposed Method

We propose *MonoPop3D*, a hybrid computational photography pipeline designed to transform a single, unconstrained RGB image into a volumetric 3D animation. Unlike photogrammetry or Neural Radiance Fields (NeRF) which require multi-view constraints, our system relies on a single view. To resolve the inherent ambiguity of single-view reconstruction—specifically for complex, fuzzy subjects like pets—we integrate deep learning priors for semantic understanding with test-time mathematical optimization for visual refinement.

The pipeline is composed of five distinct stages: (1) Interactive Instance Segmentation, (2) Boundary Refinement via Alpha Matting, (3) Deep Generative Inpainting, (4) Volumetric Warping, and (5) Photometric Enhancement. The architecture is visualized in Figure 1.

3.1 Interactive Segmentation with Topology Repair

The first challenge in single-view reconstruction is accurately isolating the foreground subject. Automated semantic segmentation models (e.g., DeepLabV3, Mask R-CNN) typically operate on a fixed set of class labels. While effective for generic categories, they lack the granularity to handle specific user intent, such as separating a pet from a cluttered background or handling occlusion (e.g., a dog's feet hidden behind a bowl).

To address this, we employ the **Segment Anything Model (SAM)** [2]. SAM is a Vision Transformer (ViT) based model trained on the SA-1B dataset (11 million images), enabling it to zero-shot generalize to arbitrary objects. We utilize SAM's promptable interface to implement an **Additive Interactive Segmentation** scheme. Let I be the input image. The user provides a set of click coordinates $\mathcal{C} = \{p_1, p_2, ..., p_N\}$. For each point p_i, SAM predicts a binary mask $M_i \in \{0,1\}^{H \times W}$. The raw cumulative mask is computed as the logical disjunction:

$$M_{raw} = \bigcup_{i=1}^{N} M_i \tag{1}$$

A common failure mode in interactive segmentation is the "cracked limb" artifact, where distinct clicks on a single object (e.g., torso and leg) result in disjoint mask islands separated by thin boundaries. To ensure the 3D object remains topologically connected, we apply a morphological optimization step called **Gap Stitching**. We define a structuring element K of size 7×7 and perform a morphological closing operation:

$$M_{final} = (M_{raw} \oplus K) \ominus K \tag{2}$$

where $\oplus$ denotes dilation and $\ominus$ denotes erosion. This operation bridges small gaps between clicked regions without significantly expanding the object's silhouette.

3.2 Boundary Optimization: Alpha Matting

Standard segmentation pipelines output binary masks, which classify pixels as either fully foreground or fully background. This binary assumption fails catastrophically for "fuzzy" subjects such as pets, teddy bears, or hair. A binary cut produces jagged, artificial edges (the "aliasing" artifact) and eliminates the fine high-frequency details that constitute fur.

To recover these details, we introduce an **Alpha Matting** module based on the **Guided Filter** [4]. Unlike deep learning matting methods that require a precise trimap input, the Guided Filter is a computationally efficient, edge-preserving smoothing operator that solves a local linear model.

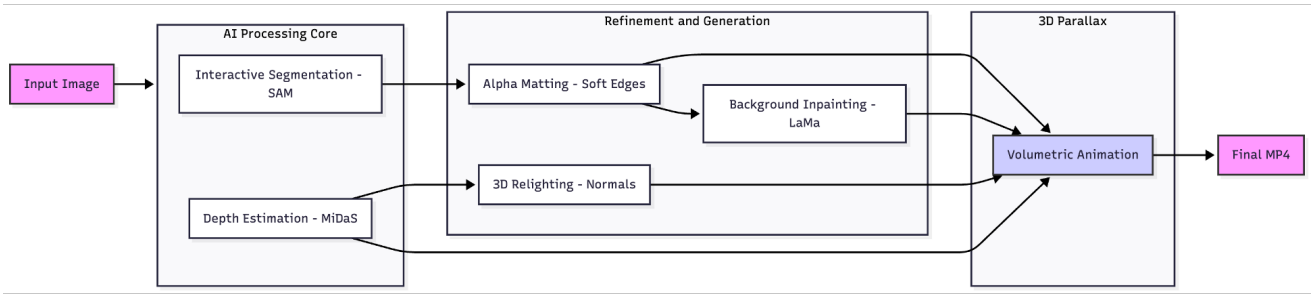

Fig. 1. The MonoPop3D Architecture. We combine the semantic reasoning of large pre-trained models (SAM, LaMa, MiDaS) with low-level signal processing (Guided Filtering, Morphological stitching) to achieve high-fidelity 3D parallax effects from a single image.

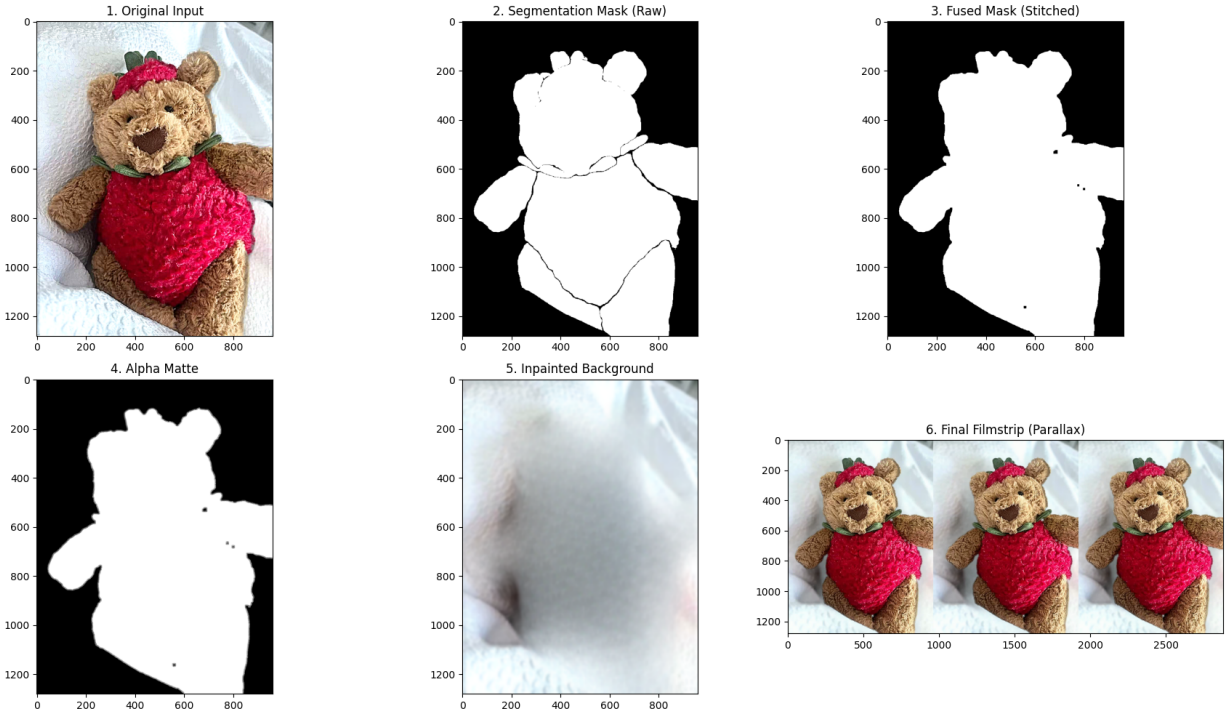

Fig. 2. Intermediate pipeline outputs. From figure 1 to figure 6: Original Image, Binary SAM Mask, fused SAM Mask, Refined Alpha Matte (showing fuzzy edges), the Hallucinated Background generated by LaMa, and the final filmstrip generated.

We first automatically generate a trimap T from our binary mask M_{final}. We define three regions: definitive foreground Ω_F, definitive background Ω_B, and an unknown transition region Ω_U:

$$T(x) = \begin{cases} 1 & \text{if } x \in \text{Erode}(M_{final}, K_{small}) \\ 0 & \text{if } x \notin \text{Dilate}(M_{final}, K_{small}) \\ 0.5 & \text{otherwise (Unknown)} \end{cases} \quad (3)$$

The Guided Filter then computes the refined alpha matte α by using the original RGB image I as a guidance map. The assumption is that sharp edges in the alpha channel should align with sharp edges in the color channel. The output α_i at pixel i is modeled as a linear transform of the guidance image intensity I_i within a local window ω_k:

$$\alpha_i = a_k I_i + b_k, \quad \forall i \in \omega_k \quad (4)$$

By minimizing the squared difference between the modeled α and the input trimap T, we recover fractional alpha values ($\alpha \in [0, 1]$) in the unknown region. This effectively recovers individual strands of fur and hair, allowing the background to "bleed through" the edges of the pet naturally during parallax movement.

3.3 Background Hallucination via LaMa

Isolating the foreground leaves a void (hole) in the background layer. Traditional inpainting techniques (e.g., Navier-Stokes, Telea) utilize diffusion-based propagation, which fills holes by blurring inward from the boundary. While fast, this destroys structural consistency, leaving a visible "ghost" behind the moving object.

We instead employ **LaMa (Large Mask Inpainting)** [3]. The rationale for choosing LaMa lies in its use of **Fast Fourier Convolutions (FFCs)**. Standard Convolutional Neural Networks (CNNs) suffer from limited receptive fields, making them poor at connecting distant structures across large holes (e.g., the entire body of a bear). FFCs operate in the spectral domain, covering the entire image globally. This allows the model to understand global context—such

Fig. 3. Demonstrate on Relighting and Bokeh Effects. Image A and B are image without relighting vs image with relighting. Image C and D are image without Bokeh and image with Bokeh.

as the repetitive pattern of a floor or the gradient of a wall—and hallucinate semantically consistent structures to fill the void. This ensures that as the foreground object moves, the revealed background appears physically plausible.

3.4 Volumetric 3D Warping

A major limitation of existing "2.5D" photo animation tools is the planar assumption: the foreground is treated as a flat billboard ($z = \text{const}$). This results in a "cardboard cutout" effect where the object slides laterally but exhibits no internal parallax.

To simulate true 3D volume, we implement a non-linear displacement mapping driven by **Monocular Depth Estimation**. We utilize **MiDaS** [7], a transformer-based depth estimator trained on diverse datasets, to predict a dense relative depth map D. We define the horizontal pixel shift Δu as a function of the animation time t and the normalized depth $\hat{D}(x, y)$. Crucially, we introduce a non-linear convexity term γ to model the object's volume:

$$\Delta u(x, y) = S_{base} \cdot \sin(\omega t) \cdot (1 + \lambda_{vol} \cdot \hat{D}(x, y)^{\gamma}) \quad (5)$$

Here, S_{base} is the global layer shift and λ_{vol} is a user-controlled volume strength parameter. By raising the normalized depth to a power $\gamma > 1$ (typically $\gamma = 1.5$), we exaggerate the curvature of the object's center (e.g., the nose) relative to its periphery (e.g., the ears). This differential motion mimics the optical flow of a rotating 3D sphere, creating a convincing illusion of volume without explicit mesh reconstruction. The full pipeline's immediate results are visualized in Figure 2.

3.5 Novelty: Photometric Enhancement

To further integrate the synthesized geometry into the scene, we introduce two rendering enhancements that exploit the estimated depth map for stylistic realism. The effect of both modules is visualized in Figure 3.

TABLE 1
Quantitative comparison between baseline processing and the full MonoPop3D pipeline.

Metric	Baseline	Ours	Change
Gradient Mean (Texture)	438.46	**626.69**	+42.9%
High-Freq Energy (BG)	3.91×10^7	$\mathbf{2.56 \times 10^7}$	-34.4%
Inpainting PSNR	38.61 dB	**38.77 dB**	+0.16
Max Gradient (Edge)	4.47	**1.21**	-73.0%

3.5.1 *Normal-Based Relighting*

A common artifact in parallax animations is static lighting; as an object rotates, its specular highlights should shift, but in a 2D image, they remain frozen. To correct this, we synthesize dynamic lighting. We first compute the surface normals $\mathbf{n}$ by taking the gradient of the depth map D:

$$\mathbf{n}(x, y) = \text{Normalize}\left(-\frac{\partial D}{\partial x}, -\frac{\partial D}{\partial y}, 1\right) \quad (6)$$

We assume a virtual directional light source $\mathbf{l}$ and apply a Lambertian shading model. The relit pixel intensity I'_{rgb} is given by:

$$I'_{rgb} = I_{rgb} \cdot (1 + \beta \cdot (\mathbf{n} \cdot \mathbf{l} - 0.5)) \quad (7)$$

where β controls the relighting intensity. This introduces new highlights on surfaces facing the virtual light and shadows on opposing surfaces, reinforcing the 3D shape perception of the fuzzy subject.

3.5.2 *Depth-Dependent Bokeh*

Macro photography of pets and toys typically exhibits a shallow depth of field. To replicate this, we implement a synthetic Bokeh effect. Since our pipeline strictly separates foreground and background layers, we apply a Gaussian blur kernel G_σ selectively to the background layer B, where the kernel size σ mimics the circle of confusion for a lens focused on the foreground:

$$B'_{bokeh} = B * G_\sigma \quad (8)$$

This low-pass filtering suppresses high-frequency noise in the hallucinated background, directing the viewer's visual attention to the sharp, volumetrically warped foreground object.

4 Experimental Results

We evaluate MonoPop3D by performing a systematic ablation study. The evaluation focuses on the contributions of four key modules: Boundary Refinement (Alpha Matting), Background Hallucination (Inpainting), Volumetric Warping, and Photometric Enhancement. We provide both a quantitative comparison of critical visual metrics and a qualitative analysis of the resulting visual improvements.

4.1 Quantitative Evaluation

To assess the numerical characteristics of our approach, we define four domain-specific metrics to objectively quantify the performance gains, particularly in addressing the challenges associated with single-view pop-up generation for complex, non-rigid subjects (e.g., subjects with fur).

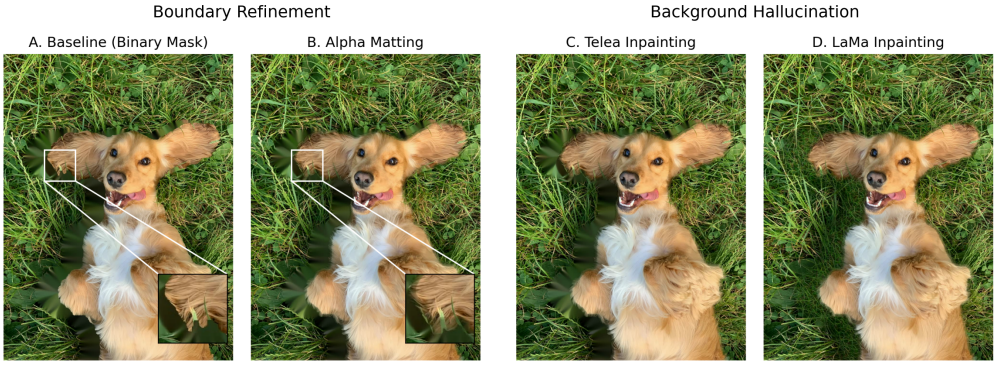

Fig. 4. Refinement ablation for the foreground and background layers. A: Baseline (Binary Mask). B: Alpha Matting. C: Telea Inpainting. D: LaMa Inpainting. Zoomed insets highlight differences in boundary quality.

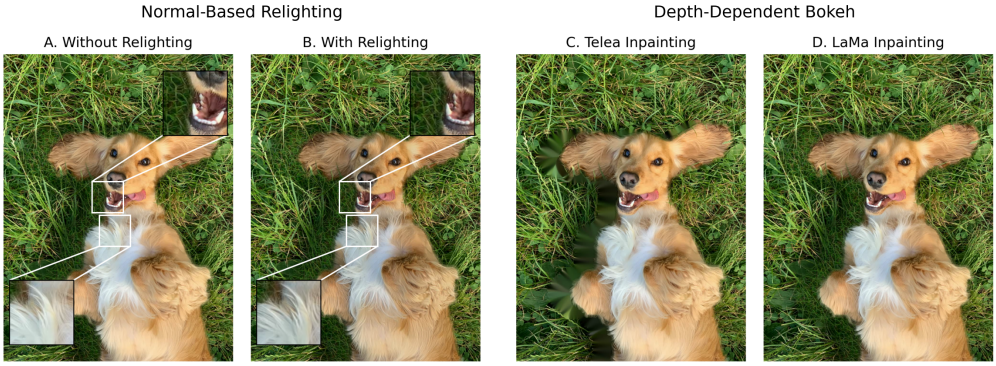

Fig. 5. Photometric ablation for relighting and background defocus. A: Without Relighting. B: With Relighting. C: Without Depth-Dependent Bokeh. D: With Depth-Dependent Bokeh. Zoomed insets highlight changes in shading.

The results presented in Table 1 are obtained by randomly sampling images from the dataset and averaging the metrics across the samples. The **Baseline** configuration against which we ablate represents a standard, minimal 3D pop-up system: no Alpha Matting (relying on a hard segmentation mask), Telea Inpainting, no Relighting, and no Bokeh effect.

Gradient Mean (Texture) provides a global measure of local texture contrast by averaging the magnitude of image gradients. As reported in Table 1, this value increases by **+42.9%** after applying normal-based relighting, indicating the added surface variation introduced by synthetic shading.

High-Freq Energy (BG) quantifies the amount of fine-scale structure present in the background layer and it decreases by **34.4%** with depth-dependent bokeh, reflecting its role in suppressing background clutter and directing visual focus toward the foreground subject.

Inpainting PSNR quantifies the fidelity of recovered content by comparing an inpainted region against the corresponding region in the original image. For this evaluation, we remove a patch from the input image, apply either the Telea baseline or LaMa to fill it, and compute PSNR against the original. As reported in Table 1, LaMa achieves a PSNR of 38.77 dB compared to 38.61 dB for Telea. It shows only a small improvement that reflects LaMa's slightly stronger ability to preserve structural coherence in the filled region.

Max Gradient (Edge) measures the strongest gradient magnitude along the foreground silhouette and, therefore, serves as an indicator of boundary sharpness. Using a binary mask in the baseline configuration yields a relatively high value of 4.47, reflecting the hard, aliased edges characteristic of unrefined cutouts. Incorporating Guided Filter Alpha Matting reduces this value by **73.0%** to 1.21, demonstrating the importance of matting for recovering fine, semi-transparent details such as fur and achieving smoother, more natural boundary transitions.

4.2 Qualitative Evaluation

To complement the quantitative results, we visualize the effect of each module in isolation through targeted ablation studies. These examples illustrate how our design choices improve boundary quality, background reconstruction, photometric realism, and perceived volumetric motion. The corresponding results are shown in Figures 4, 5, and 6.

Boundary Refinement. Figure 4 (A–B) compares the baseline cutout produced with a binary foreground mask against our Guided Filter Alpha Matting. The baseline exhibits jagged, aliased edges that create a visible "cutout" artifact, particularly around the fur. Alpha matting restores fine, semi-transparent hair structures and produces a smooth transition into the background, consistent with the reduced Max Gradient (Edge) observed in the quantitative results.

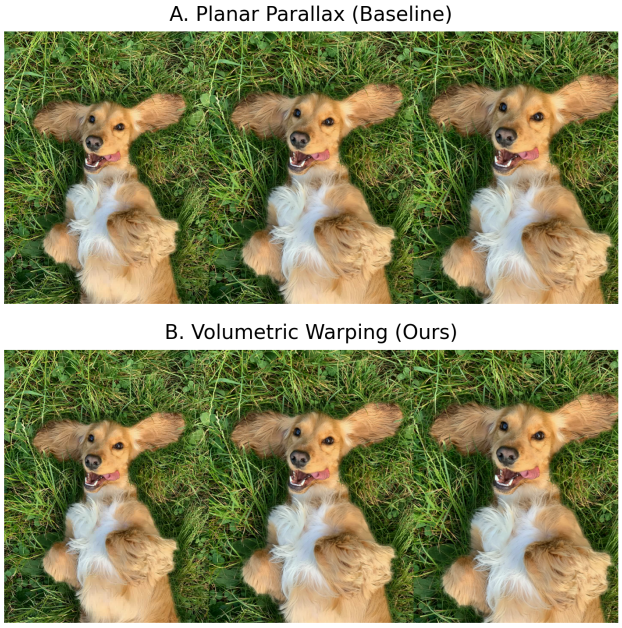

Fig. 6. Motion ablation comparing planar parallax with our non-linear volumetric warping. A: Planar Parallax (Baseline). B: Volumetric Warping (Ours). The volumetric motion exhibits internal depth variation and curvature cues, in contrast to the rigid sliding characteristic of planar motion.

Background Hallucination. Figure 4 (C–D) visualizes the inpainted background produced by Telea and LaMa. In the grassland setting used for our ablations, Telea often introduces blurry patches and locally inconsistent texture transitions within regions previously occluded by the foreground. In contrast, LaMa preserves the fine, directional structure of the surrounding grass, yielding background completions that remain coherent and clearly observable. Although the PSNR improvement observed in the quantitative evaluation is numerically small, the qualitative differences are substantial: LaMa reconstructs more plausible global structure and sharper local details, as illustrated in the figure.

Normal-Based Relighting. Figure 5 (A–B) compares the system with and without relighting. Without relighting, the subject appears uniformly illuminated, lacking the curvature cues needed to convey shape. Adding normal-based shading introduces subtle highlight and shadow variations that strengthen the perception of surface geometry. In the zoomed insets, these differences become especially apparent: fine-scale structures such as the teeth and fur exhibit more convincing depth and curvature, giving the foreground a stronger sense of three-dimensionality and overall visual realism. These perceptual improvements are consistent with the increase in Gradient Mean (Texture) reported quantitatively.

Depth-Dependent Bokeh. Figure 5 (C–D) shows the effect of applying depth-dependent blur to the background. The baseline retains sharp, distracting background detail, while our bokeh module produces a more natural separation between foreground and background. This perceptual improvement corresponds to the reduction in High-Freq Energy (BG) reported quantitatively.

Volumetric Warping (Motion). Finally, Figure 6 compares motion synthesized using planar parallax with our non-linear volumetric warping. The planar baseline (A) produces rigid, uniform lateral movement, resulting in the well-known "cardboard cutout" effect. In contrast, volumetric warping (B) applies depth-dependent displacements that create differential motion across the subject. When paying close attention to the dog's facial features, such as the relative position of the eyes and nose with respect to the edges of the face, the effect becomes especially noticeable: the face appears to rotate subtly rather than slide rigidly. This induces a stronger sense of camera panning or object rotation, yielding a more convincing three-dimensional motion consistent with the intended volumetric interpretation of the scene.

5 Discussion

We have presented MonoPop3D, a pipeline that successfully converts single RGB images into volumetric 3D pop-up animations by synergizing deep learning priors with test-time optimization. Our results demonstrate that handling fuzzy boundaries via alpha matting and simulating volume via non-linear warping significantly enhances perceptual realism compared to standard planar parallax methods.

5.1 Limitations

Despite the improved visual fidelity, our approach remains constrained by the fundamental limitations of single-view reconstruction:

- **Deep Occlusion:** Our method cannot recover geometry that is completely occluded in the input image (e.g., the back of the object or feet hidden behind obstacles). While LaMa effectively hallucinates background textures, it cannot hallucinate complex foreground structures that were never seen. The resulting mesh is essentially a "2.5D" relief rather than a full 3D model, limiting the camera trajectory to frontal hemispheres ($\pm 30°$).
- **Depth Ambiguity:** Monocular depth estimators (MiDaS) provide relative, not metric, depth. In scenes with unusual perspectives or lack of texture, the estimated depth map may be flattened or distorted, leading to incorrect volumetric warping (e.g., a flat face appearing concave).
- **Thin Structures:** Although alpha matting recovers fur well, extremely thin or semi-transparent structures (e.g., insect wings, glass) pose a challenge for both segmentation (SAM) and depth estimation, often resulting in artifacts where the background bleeds into the foreground.

5.2 Future Work

Several avenues exist to extend this research toward fully immersive 3D experiences:

- **Generative 3D Completion:** To solve the occlusion problem, future work could integrate text-to-3D generative models (e.g., DreamFusion or Shap-E). Instead of just inpainting the background, we could

use the segmented mask as a condition to generate a complete, textured 3D mesh of the object, allowing for 360° rotation.

- **Multi-View Fusion:** For users who can provide 2-3 images (e.g., a stereo pair or a short video burst), the pipeline could be adapted to utilize multi-view stereo (MVS) or sparse NeRF optimization. This would replace the "hallucinated" depth with metric depth, correcting geometric distortions.
- **Neural Rendering:** Replacing the explicit mesh/layer rendering with a neural renderer (e.g., a lightweight MLP) could allow for view-dependent effects, such as realistic fur shading that changes with light direction, surpassing our current normal-based relighting approximation.

6 Conclusion

MonoPop3D represents a step toward democratizing 3D content creation by removing the need for multi-view capture or professional 3D modeling tools. Our lightweight pipeline combines pretrained deep models with classical image processing techniques, illustrating the strength of hybrid neuro-symbolic systems in which AI provides the initial estimate and mathematical optimization refines the result. Through the integration of matting, inpainting, relighting, and volumetric warping, MonoPop3D produces smoother boundaries, more coherent background completions, realistic shading cues, and convincing depth-dependent motion from a single image. The results demonstrate that carefully designed photometric and geometric refinements can substantially enhance the realism of 2.5D animations while remaining computationally efficient. Future extensions include handling articulated subjects, incorporating learned volumetric priors, and deploying the framework in interactive media and creative applications.

Acknowledgments

The authors would like to thank Dr. David Lindell for teaching CSC2529 and for providing insightful guidance throughout the course. We also extend our gratitude to our team's pets: Paopao, Coral, and Pearl, for inspiring this project and for generously contributing their photographs for our experiments.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: http://arxiv.org/abs/2003.08934v2

[2] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[3] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2149–2159.

[4] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[5] J. Shade, S. Gortler, L.-W. He, and R. Szeliski, "Layered depth images," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1998, pp. 231–242.

[6] M.-L. Shih, S.-Y. Su, J. Kopf, and J.-B. Huang, "3d photography using context-aware layered depth inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[7] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, 2020.