# $\alpha$Depth: Learning Scene-Level Soft Boundary Decomposition for Stereo Conversion

**Xiang Zhang**[1,2], **Yang Zhang**[2], **Lukas Mehl**[2],
**Karlis Martins Briedis**[2], **Markus Gross**[1,2], **Christopher Schroers**[2]
[1]ETH Zürich, [2]DisneyResearch|Studios

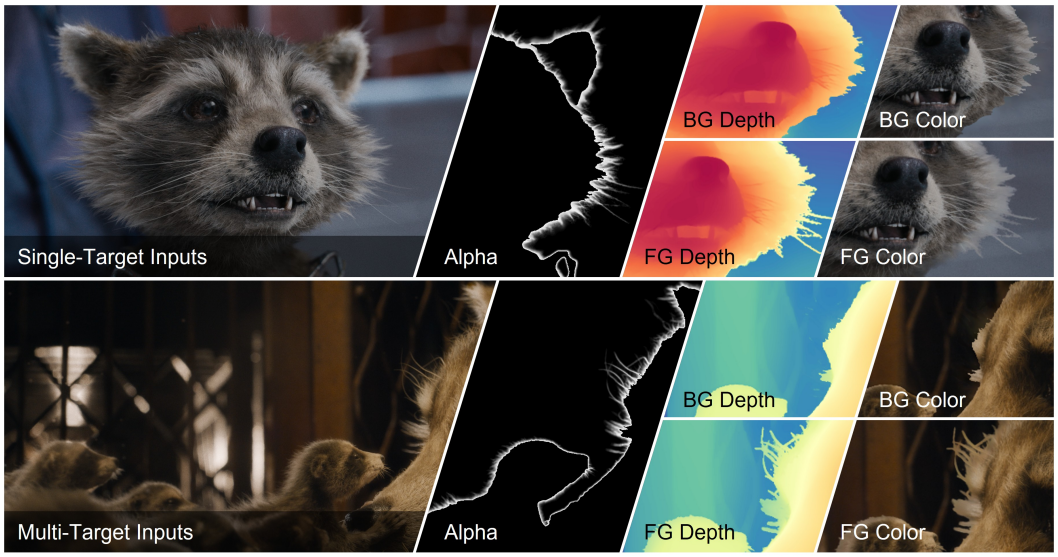



Figure 1: **Layered $\alpha$Depth Representation**. We introduce $\alpha$Depth to decompose soft boundaries (*e.g.*, hair, thin structures, and defocus blur) for high-fidelity stereo conversion. Given an image and its depth map as inputs, our approach estimates layered information, *i.e.*, alpha, foreground/background (FG/BG) colors and depths, at *local soft boundaries* (see non-zero alpha regions), enabling scene-level inference of multiple/overlapping targets in a single forward pass without user intervention.

## Abstract

Accurately modeling soft boundaries, *e.g.*, hair and defocus blur, is a fundamental challenge in stereo conversion due to the ambiguous blending of foreground and background. Existing depth models primarily predict single-layer depth, leading to ambiguity in depth correspondence at soft boundaries. While matting techniques can capture opacity for layered modeling, they often struggle in complex scenes with multiple targets and usually require user intervention. This paper introduces **$\alpha$Depth**, a layered representation that decomposes soft boundaries for high-fidelity stereo conversion. Specifically, we first resolve mixed color and depth ambiguity by estimating layered color and depth values at soft boundaries. Considering complex multi-target scenes, we design a Circular Alpha Representation (CAR) that shifts the paradigm from global target extraction to local boundary decomposition. Unlike prior matting methods restricted to a single foreground/background, CAR enables efficient scene-level inference without manual guidance. Extensive evaluations demonstrate that $\alpha$Depth achieves state-of-the-art performance in stereo conversion, eliminating background bleeding and structural distortions at soft boundaries.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

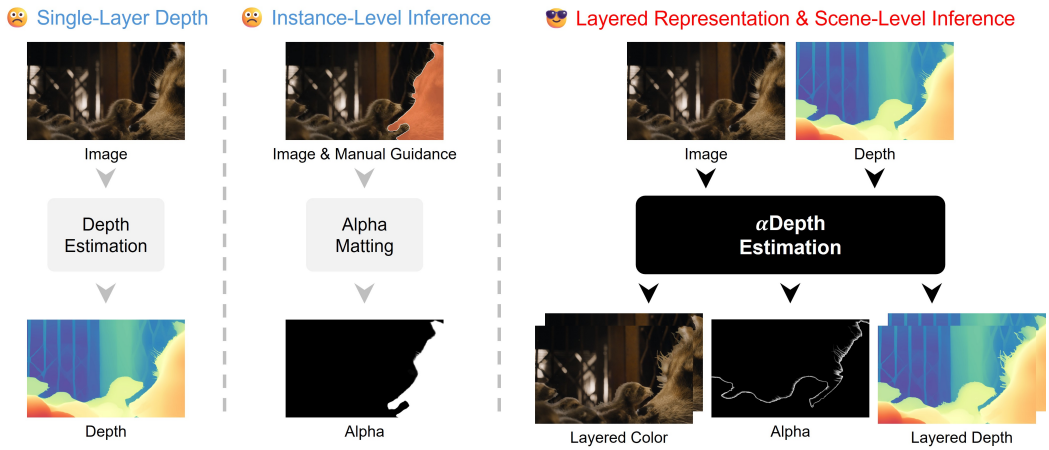



Figure 2: **Comparison with existing paradigms.** Depth estimation models typically assign a single depth value per pixel, struggling with mixed colors at soft boundaries and suffering from depth ambiguity. While conventional matting approaches extract instance-level soft boundaries, they usually require manual guidance (*e.g.*, trimaps). In contrast, our layered $\alpha$Depth representation enables automatic scene-level decomposition. Given an image and its depth map, $\alpha$Depth explicitly decomposes soft boundaries into layered color, alpha, and layered depth in a single forward pass.

## 1 Introduction

By lifting 2D images into 3D content, stereo conversion is critical for various immersive applications, including Virtual/Augmented Reality (VR/AR) and movie production [26, 62, 58, 8]. Recent methods have demonstrated remarkable progress in stereo conversion by leveraging foundation diffusion models to synthesize realistic novel views from monocular inputs [57, 43, 39]. For instance, Eye2eye [10] effectively utilizes diffusion priors to simulate complex view-dependent effects such as specular reflections in stereo conversion. Approaches like SplatDiff [61] and Elastic3D [27] have focused on improving the overall visual fidelity and structural consistency of the generated stereo pairs. Despite these advances, generating high-quality stereo content from monocular inputs remains challenging, particularly when dealing with intricate details in complex scenes.

A key challenge in stereo conversion is the accurate handling of *soft boundaries*, which naturally occur across diverse subjects (*e.g.*, humans, animals, or computer-generated characters as shown in Fig. 1) and camera effects (*e.g.*, defocus blur). At soft boundaries, foreground and background colors inherently mix within a single pixel, creating regions of partial transparency and resulting in ambiguous depth correspondence. Consequently, previous stereo conversion methods often struggle to recover soft boundary details at high fidelity and produce stereo results with visual artifacts, such as halo effects, background bleeding, or unnatural floating textures around object silhouettes [62].

To achieve high-quality soft boundary recovery in stereo conversion, two fundamental challenges exist: **(i) Mixed Foreground and Background**: Many stereo conversion pipelines rely on monocular depth estimation for view synthesis [57, 40, 61, 63]. However, conventional depth estimators typically assign only a single depth value per pixel [51, 2, 60, 11], failing to model the layered characteristics of soft boundaries (Fig. 2). Although the recent work HairGuard [62] better captures soft boundary structures via depth refinement, it still suffers from depth ambiguity due to the single-layer depth representation. **(ii) Complex Scenes with Multiple Targets**: Alpha matting techniques can extract alpha mattes for soft boundary decomposition [55, 15, 13, 53]. However, existing matting methods generally rely on manual guidance, such as trimaps, visual prompts (*e.g.*, points or boxes), and segmentation masks, for instance-level inference (Fig. 2). Consequently, these techniques necessitate user intervention for each target or repeated forward passes, rendering them impractical for automated stereo conversion pipelines when handling complex multi-target scenes. While auxiliary-free matting approaches like GVM [9] are emerging, they are typically designed for specific target categories (*e.g.*, humans or animals) and thus struggle to generalize to the diverse types of soft boundaries present in complex stereo conversion scenarios (*e.g.*, defocus blur in movie shots).

We introduce $\alpha$**Depth**, a novel layered representation designed to explicitly decompose local foreground and background at soft boundaries for high-fidelity stereo conversion, as shown in Fig. 2. To address the challenge of layer mixing, our $\alpha$Depth jointly estimates foreground and background information to model soft boundary regions. This explicitly disentangles the mixed colors and resolves depth ambiguity by allocating distinct depth and color values to the overlapping layers at soft boundaries. To enable efficient scene-level inference in complex scenarios, we introduce the Circular Alpha Representation (CAR). Unlike vanilla alpha representations that rely on global foreground and background definitions, CAR shifts the focus to modeling opacity exclusively at local soft boundaries. It treats all opaque regions (whether foreground or background) as a single unified class (*e.g.*, see the black regions in Fig. 2). Furthermore, while resolving all occlusions might theoretically demand numerous layers, we observe that *a two-layer formulation is effective in locally separating foreground and background at soft boundaries, even in complex scenes with many layers.* This localized approach allows $\alpha$Depth to automatically decompose multiple overlapping targets across the entire scene in a single forward pass without user intervention. Leveraging the estimated $\alpha$Depth representation, we perform layered warping to synthesize initial novel views, which are subsequently refined by off-the-shelf inpainting models to generate high-fidelity stereo pairs. Finally, we introduce an efficient training data curation method that leverages existing image datasets and matting datasets to construct training pairs for $\alpha$Depth estimation. In summary, our main contributions are:

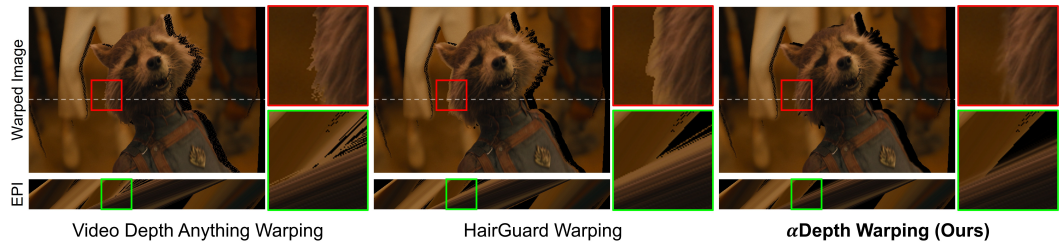
- We propose a novel layered $\alpha$Depth representation that explicitly disentangles mixed colors and resolves depth ambiguity at soft boundaries for high-fidelity stereo conversion.
- We design the Circular Alpha Representation (CAR) to model local soft boundary transitions rather than extracting global foregrounds, enabling efficient scene-level inference in a single forward pass without user intervention.

## 2 Related Work

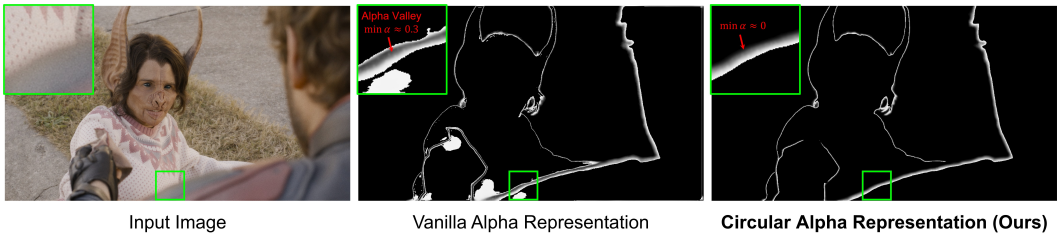
**Stereo Conversion.** Stereo conversion aims to synthesize high-quality stereo pairs from monocular inputs [48, 10, 4], a crucial technique for immersive media applications [26, 57]. Recent methods often leverage generative foundation models: image-based approaches [43, 57] utilize diffusion priors, *e.g.*, Stable Diffusion [37], to generate realistic stereo views, while video-based methods [63, 10, 40] design spatio-temporal mechanisms, *e.g.*, tiled diffusion strategy [63] and global spatial attention [40], for improved temporal coherence. To mitigate the texture hallucinations and geometric distortions inherent to diffusion models, several works incorporate explicit guidance mechanisms, such as texture bridge [61], guided decoding [27], and color fuser [62], for enhanced stereo conversion fidelity. However, due to the mixed foreground and background, existing approaches often struggle with soft boundaries and produce stereo results with visual artifacts like background bleeding.

**Depth Estimation.** Monocular depth estimation predicts dense scene geometry from a single image, serving as a fundamental pillar for 3D vision tasks like stereo conversion [35, 51, 57, 61]. To achieve robust zero-shot generalization, a dominant paradigm is to train depth models on large-scale datasets [19, 56, 35, 51, 44, 33]. For instance, MiDaS [36] proposes a family of depth losses to enable model training on diverse datasets. Recently, a variety of techniques have been developed to combine real-world and synthetic datasets for enhanced detail extraction, such as teacher-student distillation [52, 20], edge-guided loss [32], real data refinement [45], and training protocols for fine boundary preservation [2]. Alternatively, harnessing geometric priors from pre-trained generative models has emerged as a compelling trajectory. Methods like Marigold [11], BetterDepth [60], and Pixel-Perfect Depth [49] cast depth estimation as an iterative denoising process to achieve remarkable detail recovery. Nevertheless, existing depth estimators generally assign a single depth value per pixel. This single-layer representation fundamentally fails at handling soft boundaries, suffering from depth ambiguity where foreground and background overlap.

**Alpha Matting.** Alpha matting estimates the opacity of foreground targets, *i.e.*, alpha mattes, to extract foreground objects from their backgrounds [50, 55, 53], which potentially benefits soft boundary modeling in stereo conversion. To resolve semantic ambiguities, previous matting models usually rely on manual guidance, *e.g.*, trimaps, to explicitly define foreground and background [50, 31, 55]. Recent approaches have developed more flexible guidance, *e.g.*, segmentation masks or visual prompts, to facilitate matting for video sequences [54, 53]. To reduce human effort, auxiliary-free methods, such as MODNet [12], RVM [22], and GVM [9], leverage pre-trained model priors,




(a) Visual comparison of warping performance and Epipolar Plane Images (EPIs)




(b) Alpha valley issue (Alpha is shown only at soft regions, *i.e.*, $\alpha \in [0.02, 0.98]$, for comparison)

Figure 3: **Challenges of soft boundary recovery in stereo conversion.** (a) We evaluate warping performance via Epipolar Plane Images (EPIs) extracted along the gray dashed line under uniform rightward camera motion. Direct warping with Video Depth Anything [3] struggles with depth ambiguity at soft boundaries, causing broken edges and flying pixels. Although HairGuard [62] captures better details, its single-layer depth representation often results in background bleeding and aliasing (*e.g.*, see EPIs). By contrast, our layered representation effectively models soft boundaries to achieve superior warping and multi-view consistency. (b) Vanilla alpha representations often suffer from "alpha valleys" (*i.e.*, alpha estimation errors at intersecting boundaries) due to their reliance on explicit global foreground and background definitions. By modeling local transitions rather than global separation, our Circular Alpha Representation (CAR) robustly handles overlapping targets to produce accurate alpha values.

*e.g.*, diffusion priors, or temporal consistency to infer alpha mattes directly from input images. However, both paradigms exhibit critical limitations in complex scenes. Guidance-based methods necessitate user intervention or repeated forward passes for each subject, rendering them impractical for integration into automated stereo conversion pipelines. Conversely, auxiliary-free methods are generally restricted to specific semantic categories (*e.g.*, humans) and struggle to generalize to diverse soft boundaries (*e.g.*, defocus blur). To overcome these bottlenecks, we introduce the Circular Alpha Representation (CAR). By shifting the paradigm from global foreground extraction to local soft boundary decomposition, CAR enables $\alpha$Depth to automatically infer opacity across complex scenes in a single forward pass.

## 3 Method

We first analyze the main challenges of soft boundary recovery in stereo conversion (Sec. 3.1), and then propose the layered $\alpha$Depth representation for efficient soft boundary decomposition (Sec. 3.2).

### 3.1 Problem Analysis

Stereo conversion in real-world scenarios frequently encounters complex scenes characterized by multiple overlapping targets and intricate soft boundaries (*e.g.*, hair, fur, and defocus blur). At these soft boundaries, the observed color $I$ is inherently a mixture of the foreground color $I_{\mathrm{FG}}$ and the background color $I_{\mathrm{BG}}$, modulated by an opacity value $\alpha \in [0, 1]$. Following alpha matting [55, 53], the color blending at soft boundaries can be defined as:

$$I = \alpha I_{\mathrm{FG}} + (1 - \alpha) I_{\mathrm{BG}}. \tag{1}$$

Mixed colors at soft boundaries, coupled with multiple targets, introduce significant challenges:

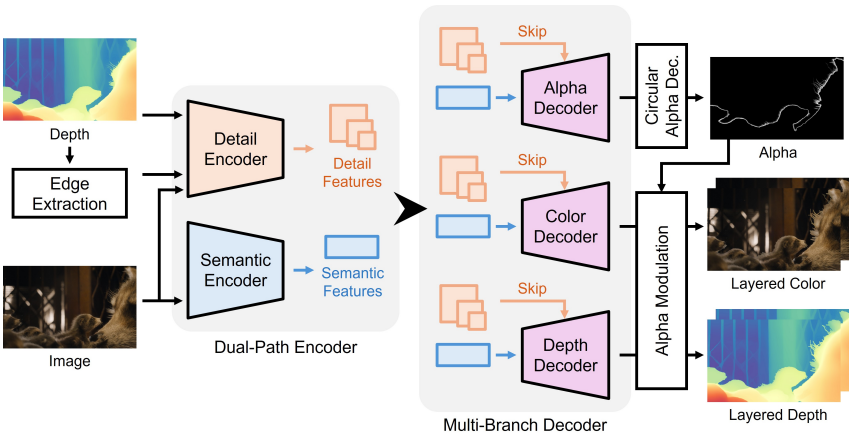



Figure 4: $\alpha$**Depth estimation pipeline.** Given an image and its corresponding depth map (*e.g.*, from a pre-trained depth model), we employ a dual-path encoder to extract both semantic and detail features. A multi-branch decoder then processes these features for task-specific predictions. Finally, we apply circular alpha decoding to generate the estimated alpha map, which subsequently modulates and constrains the layered color and depth predictions on soft boundary regions.

**Depth Ambiguity.** Based on Eq. (1), pixels within a soft boundary mix color contributions from both foreground and background layers. However, most monocular depth estimators predict only a single depth value per pixel [51, 11, 44, 49]. This single-layer representation inherently results in depth ambiguity at soft boundaries. Consequently, view transformation techniques like softmax splatting [29] struggle to project these pixels accurately, leading to broken boundaries and severe flying pixels (Fig. 3a). Although recent methods like HairGuard [62] capture finer boundary details, their single-layer depth often leads to background bleeding and aliasing artifacts (Fig. 3a).

**Scene-level Inference.** Boundaries at depth discontinuities are crucial for high-fidelity stereo conversion, as they govern the realistic rendering of occlusions and binocular parallax [57, 62]. However, complex scenes often feature multiple overlapping targets at varying depths (*e.g.*, Fig. 3b), and even individual targets may possess intricate soft boundaries within their own structures (*e.g.*, see alpha in the top example of Fig. 1). Existing matting methods typically rely on a global definition of foreground and background, *e.g.*, foreground defined by user inputs like trimaps or visual prompts [55, 53]. Consequently, they either require iterative user guidance for each instance or indiscriminately merge multiple targets into a single monolithic foreground, discarding vital inter-object boundary details.

Furthermore, learning a vanilla alpha representation for multiple overlapping targets often suffers from the "*alpha valley*" issue (Fig. 3b). To preserve local boundary details of overlapping foreground instances, alpha discontinuities emerge at their intersections (*e.g.*, see alpha in Fig. 5). Since neural networks are usually biased toward smooth predictions, they often struggle to resolve these sharp transitions, yielding inaccurate alpha values (Fig. 3b). These estimation errors ultimately propagate through the stereo conversion pipeline, yielding inconsistent soft boundaries in the stereo pairs.

### 3.2 $\alpha$Depth

We propose a layered $\alpha$Depth representation to address depth ambiguity and alpha valley issues for efficient soft boundary decomposition. Instead of using manual guidance like trimaps, we directly utilize image semantics and scene geometry for $\alpha$Depth estimation, enabling scene-level inference without user intervention (Fig. 4). Given an input image $I_{\text{IN}}$ and its depth map $D_{\text{IN}}$, we first extract depth gradients to facilitate soft boundary localization. Then, we employ a dual-path encoder to capture both high-level contextual cues and fine-grained structural details. Specifically, a detail encoder built upon UNet encoder [38] extracts multi-scale features to preserve rich textural and structural details. Concurrently, a semantic encoder (based on DINOv2 [30] with Depth Anything V2 weights initialization [52]) extracts deep semantic features for soft boundary reasoning. Finally, these

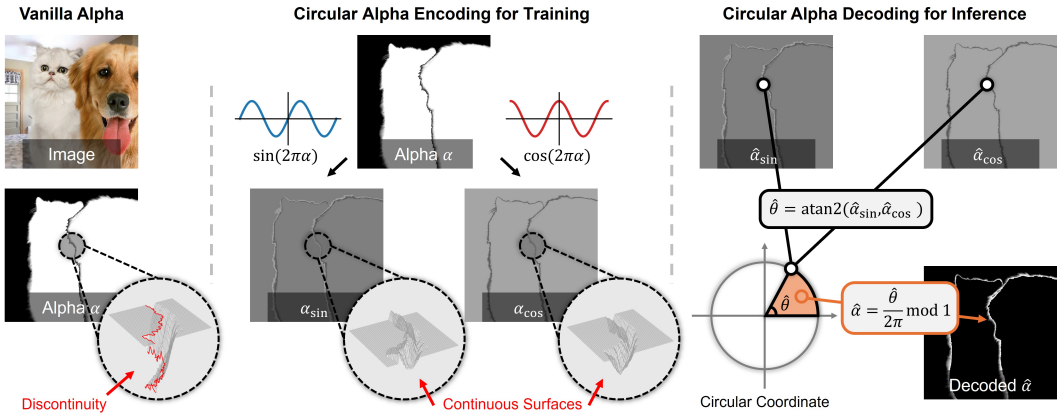



Figure 5: **Circular Alpha Representation (CAR).** The vanilla alpha representation inherently suffers from sharp discontinuities at the intersecting boundaries of multiple overlapping instances. By contrast, CAR encodes the ground-truth alpha into continuous trigonometric space during training, benefiting model optimization and eliminating alpha valley issues (Fig. 3b). During inference, the predicted trigonometric components are decoded back into an alpha matte $\hat{\alpha}$ at soft boundaries.

features are fed into a multi-branch decoder (based on UNet decoder [38]) to predict the $\alpha$Depth representation, which jointly estimates alpha, layered color, and layered depth at soft boundaries.

**Circular Alpha Representation (CAR).** Traditional matting paradigms rely on a global definition of foreground and background [55, 53]. In scenarios featuring multiple overlapping objects at varying depths, this global assignment forces sharp discontinuities at inter-object intersecting boundaries (Fig. 5), leading to the alpha valley issue (see Fig. 3b). By contrast, the proposed CAR reformulates the task by shifting from global foreground extraction to local soft boundary decomposition. We treat all opaque regions, whether foreground or background, as a single unified class and focus on estimating opacity at semi-transparent boundaries. To achieve this, we project the ground-truth alpha map $\alpha \in [0, 1]$ into a continuous trigonometric space via circular alpha encoding (Fig. 5):

$$\alpha_{\sin} = \sin(2\pi\alpha), \quad \alpha_{\cos} = \cos(2\pi\alpha). \tag{2}$$

By leveraging the periodicity of trigonometric functions, we wrap the linear alpha scale around a circle, mapping background ($\alpha = 0$) and opaque foreground ($\alpha = 1$) to the same coordinate $(0, 1)$ in the $(\alpha_{\sin}, \alpha_{\cos})$ space. This transformation collapses the discrete jump at intersecting boundaries into a continuous manifold (Fig. 5). By bridging the gap between $\alpha = 0$ and $\alpha = 1$, it eliminates boundary discontinuities and facilitates model optimization.

During inference, the alpha decoder directly estimates the continuous trigonometric components, *i.e.*, $\hat{\alpha}_{\sin}$ and $\hat{\alpha}_{\cos}$. We then apply circular alpha decoding to reconstruct alpha values $\hat{\alpha} \in [0, 1)$ via the four-quadrant inverse tangent function, *i.e.*,

$$\hat{\alpha} = \frac{\hat{\theta}}{2\pi} \bmod 1, \quad \text{where} \quad \hat{\theta} = \text{atan2}(\hat{\alpha}_{\sin}, \hat{\alpha}_{\cos}). \tag{3}$$

As shown in Fig. 5, CAR projects opaque regions into a unified class with $\hat{\alpha} = 0$ and focuses on estimating opacity for soft boundaries. By navigating this circular space, CAR circumvents alpha valley issues and enables scene-level soft boundary decomposition in a single forward pass.

**Layered Representation.** To resolve depth ambiguity and color mixing, we explicitly decouple soft boundaries into local foreground (FG) and background (BG) representations. While complex scenes may theoretically require numerous global layers to account for all occlusions, we observe that *local soft boundaries can be effectively modeled using a two-layer decomposition*, *i.e.*, locally differentiating foreground and background for each soft boundary. Thus, we adopt a two-layer representation for both color and depth at soft boundary regions. Specifically, we estimate the layered color $\bar{I}_{\text{FG}}, \bar{I}_{\text{BG}} \in \mathbb{R}^{3\times h\times w}$ and color blending weights $W^I_{\text{FG}}, W^I_{\text{BG}} \in \mathbb{R}^{h\times w}$ via the color decoder. Similarly, the depth decoder predicts the layered depth $\bar{D}_{\text{FG}}, \bar{D}_{\text{BG}} \in \mathbb{R}^{h\times w}$ and depth blending weights $W^D_{\text{FG}}, W^D_{\text{BG}} \in \mathbb{R}^{h\times w}$. In addition, we estimate soft boundary regions $\hat{S}$ by thresholding

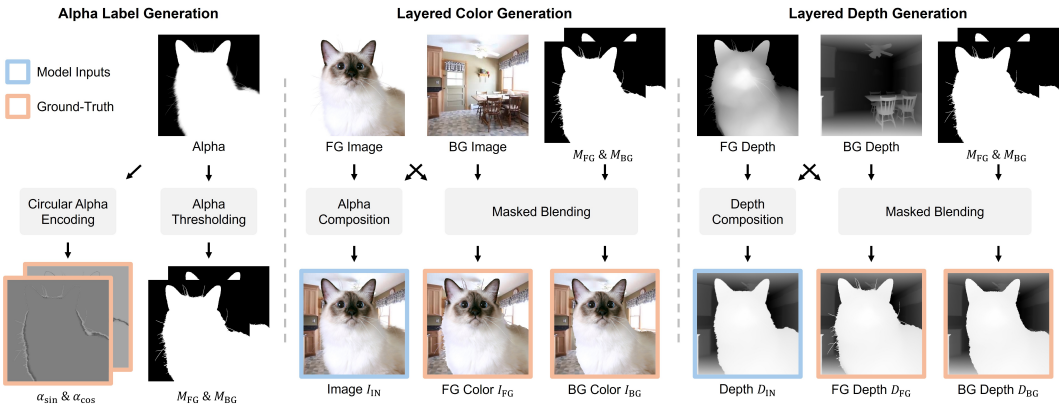



Figure 6: **Training Data Curation.** Firstly, the alpha map is processed via circular alpha encoding to yield continuous alpha labels ($\alpha_{\sin}, \alpha_{\cos}$) and thresholded to produce layered masks ($M_{\text{FG}}, M_{\text{BG}}$). In layered color/depth generation, foreground and background assets are composited to form the synthesized input image ($I_{\text{IN}}$) and depth ($D_{\text{IN}}$). Concurrently, masked blending is applied to generate ground-truth color layers ($I_{\text{FG}}, I_{\text{BG}}$) and depth layers ($D_{\text{FG}}, D_{\text{BG}}$) for soft boundary regions.

the estimated $\hat{\alpha}$, *i.e.*, $\hat{S} = \mathbb{I}(\alpha_{\text{th}} \leq \hat{\alpha} \leq 1 - \alpha_{\text{th}})$ with $\alpha_{\text{th}} = 0.02$ and $\mathbb{I}(\cdot)$ denoting the indicator function. The layered prediction is then formulated via alpha-modulated blending, *i.e.*,

$$\hat{X}_\star = W_\star^{'X} \odot \bar{X}_\star + (1 - W_\star^{'X}) \odot X_{\text{IN}}, \quad \text{with} \quad W_\star^{'X} = \hat{S} \odot (1 - W_\star^{X}), \tag{4}$$

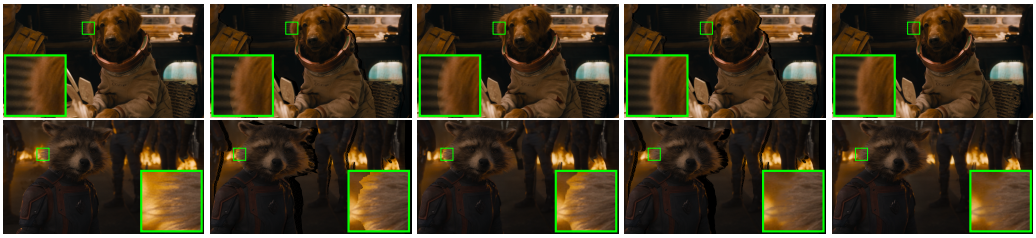
where $X \in \{I, D\}$, $\star \in \{\text{FG}, \text{BG}\}$, and $\odot$ denotes spatial element-wise multiplication. Eq. (4) restricts the layered modeling exclusively to the soft boundary regions $\hat{S}$. For opaque regions, the original input colors and depths are retained, which preserves high-fidelity textures from $I_{\text{IN}}$ and facilitates plug-and-play integration with state-of-the-art depth estimation models. We then project the estimated $\alpha$Depth representation to the target viewpoint via layered warping (see Sec. A.1 for more details). To generate the final stereo results, we employ the pretrained scene painter and color fuser from HairGuard [62] for disocclusion inpainting and texture enhancement.

**Model Training.** We propose an efficient data curation strategy to utilize existing matting datasets for $\alpha$Depth training (Fig. 6). Given a foreground RGBA image from matting datasets, we first project the ground-truth alpha matte $\alpha$ into continuous trigonometric labels $\alpha_{\sin}, \alpha_{\cos}$ via circular alpha encoding Eq. (2). We also derive the binary foreground mask $M_{\text{FG}} = \mathbb{I}(\alpha \geq \alpha_{\text{th}})$ and background mask $M_{\text{BG}} = \mathbb{I}(\alpha \geq 1 - \alpha_{\text{th}})$ for layered label generation. $M_{\text{FG}}$ captures soft boundary regions for synthesizing foreground color/depth labels, whereas $M_{\text{BG}}$ helps preserve background information at soft boundaries. For layered color generation, we composite the foreground and background (sampled from image datasets) via standard alpha blending Eq. (1) to synthesize the input image $I_{\text{IN}}$. We then apply masked blending using $M_{\text{FG}}$ and $M_{\text{BG}}$ to generate the ground-truth layered colors $I_{\text{FG}}$ and $I_{\text{BG}}$ at soft boundary regions. For layered depth generation, we synthesize the input depth map $D_{\text{IN}}$ following the depth composition protocol of HairGuard [62]. Analogously, masked blending is applied to generate ground-truth layered depths $D_{\text{FG}}$ and $D_{\text{BG}}$ (please see Sec. A.3 for more details).

The $\alpha$Depth network is trained by jointly optimizing color, depth, and alpha representations, *i.e.*,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_I + \mathcal{L}_D + \mathcal{L}_\alpha, \tag{5}$$

where the color loss $\mathcal{L}_I = \mathcal{L}(\hat{I}_{\text{FG}}, I_{\text{FG}}) + \mathcal{L}(\hat{I}_{\text{BG}}, I_{\text{BG}})$, the depth loss $\mathcal{L}_D = \mathcal{L}(\hat{D}_{\text{FG}}, D_{\text{FG}}) + \mathcal{L}(\hat{D}_{\text{BG}}, D_{\text{BG}})$, and the alpha loss $\mathcal{L}_\alpha = \mathcal{L}(\hat{\alpha}_{\sin}, \alpha_{\sin}) + \mathcal{L}(\hat{\alpha}_{\cos}, \alpha_{\cos})$. To facilitate stable multi-task learning, we apply the same loss function $\mathcal{L}(\cdot)$ across all modalities and follow a two-stage training scheme. The first-stage training focuses on recovering fine-grained details at local soft boundaries with $\mathcal{L}(\cdot)$ defined as $\mathcal{L}(\hat{X}, X) = \mathcal{L}_1(\hat{X}, X) + \mathcal{L}_{\text{m}}(S \odot \hat{X}, S \odot X)$, where $\mathcal{L}_1$ is $\ell_1$ loss, $\mathcal{L}_{\text{m}}$ denotes the matting loss from ViTMatte [55], and $S = \mathbb{I}(\alpha_{\text{th}} \leq \alpha \leq 1 - \alpha_{\text{th}})$ is the soft boundary mask. In the second stage, we apply the matting loss $\mathcal{L}_{\text{m}}$ across the entire image space for global refinement, *i.e.*, $\mathcal{L}(\hat{X}, X) = \mathcal{L}_{\text{m}}(\hat{X}, X)$.



Input Image | HairGuard Warping | HairGuard Result | **Our Warping** | **Our Result**

Figure 7: **Visual comparisons with HairGuard [62]** in warping and stereo conversion.

Table 1: **Stereo image/video conversion performance**. The best and second best results are marked.

| **Method** | **Stereo Image Conversion (Mono2Stereo)** | | | | **Stereo Video Conversion (Marvel-10K)** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | S-PSNR ↑ | S-SSIM ↑ | LPIPS ↓ | DISTS ↓ | S-PSNR ↑ | S-SSIM ↑ | LPIPS ↓ | DISTS ↓ | FVD ↓ |
| StereoDiffusion [43] | 22.95 | 0.6407 | 0.2098 | 0.0747 | 23.53 | 0.6805 | 0.2173 | 0.0694 | 18.41 |
| Mono2Stereo [57] | 24.48 | 0.6933 | 0.1990 | 0.0751 | 25.19 | 0.7309 | 0.1979 | 0.0690 | 19.81 |
| StereoCrafter [63] | 24.24 | 0.7073 | 0.2334 | 0.1198 | 25.96 | 0.7619 | 0.2287 | 0.1231 | 10.05 |
| SplatDiff [61] | 24.78 | 0.7339 | 0.1235 | 0.0439 | 26.57 | 0.8006 | 0.0831 | 0.0364 | 2.18 |
| HairGuard [62] | 25.05 | 0.7445 | 0.1212 | 0.0444 | 27.10 | 0.8239 | 0.0734 | 0.0322 | 2.14 |
| **$\alpha$Depth (Ours)** | 25.60 | 0.7554 | 0.1181 | 0.0417 | 28.46 | 0.8373 | 0.0699 | 0.0297 | 1.72 |

# 4 Experiments and Analysis

## 4.1 Experimental Settings

**Implementation**. We train the $\alpha$Depth network with AdamW optimizer [25] under $448 \times 448$ patches, batch size 32, and $1 \times 10^{-5}$ learning rate. To curate training pairs, we sample background images from RealEstate10K [64] and DL3DV-10K [23], and use foreground images from matting datasets: AM-2K [17], Distinctions-646 [34], and Composition-1K [50]. We train the model for 50 epochs per stage, which takes approximately 6 days in total on an NVIDIA RTX A6000 GPU.

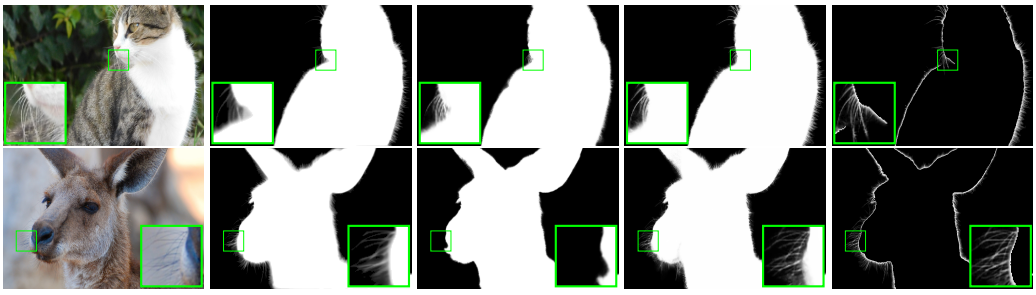
**Evaluation**. We employ the Mono2Stereo [57] and Marvel-10K [62] datasets for stereo image/video conversion. For fair comparisons, all baselines use the same depth models: Depth Anything V2 (DAv2) [52] for Mono2Stereo and Video Depth Anything (VDA) [3] for Marvel-10K. We also employ two natural image matting datasets AIM-500 [18] and P3M-10K [16] to evaluate the performance of our circular alpha representation.

## 4.2 Stereo Conversion

Tab. 1 compares $\alpha$Depth with state-of-the-art methods for stereo image/video conversion. Since soft boundaries usually occupy a small fraction of the image, we compute pixel-level metrics exclusively on soft regions $\hat{S}$ (denoted by S-PSNR and S-SSIM), alongside whole-image perceptual metrics (LPIPS [59] and DISTS [6]). Benefiting from soft boundary decomposition, our $\alpha$Depth consistently outperforms previous methods and eliminates artifacts like background bleeding (Fig. 7), yielding the best video consistency (FVD [42] in Tab. 1). Finally, Tab. 2 compares our warping performance against baselines using the original depth from VDA [3] and refined depth from HairGuard [62]. While HairGuard improves soft boundary details, its single-layer depth fails to resolve depth ambiguity (*e.g.*, mixed colors in Fig. 7). By contrast, our approach performs layered modeling on local soft boundaries and achieves the best warping performance.

Table 2: **Warping performance** on Marvel-10K [62]. Best and second best results are marked.

| **Method** | S-PSNR | S-SSIM | LPIPS | DISTS | FVD |
|---|---|---|---|---|---|
| VDA [3] | 26.14 | 0.6718 | 0.1337 | 0.0910 | 7.21 |
| VDA + HairGuard [62] | 27.90 | 0.7328 | 0.1277 | 0.0839 | 9.39 |
| **VDA + $\alpha$Depth (Ours)** | 28.68 | 0.7636 | 0.1190 | 0.0738 | 7.13 |



Input Image | GVM [9] | MatAnyone 2 [53] | ViTMatte [55] | **$\alpha$Depth (Ours)**

Figure 8: **Visual comparisons with alpha matting methods.**

Table 3: **Alpha matting performance**. The best and second best results are marked.

| Method | Guidance Type | Average Rank | AIM-500 | | | P3M-10K | | |
|---|---|---|---|---|---|---|---|---|
| | | | SAD ↓ | Grad ↓ | Conn ↓ | SAD ↓ | Grad ↓ | Conn ↓ |
| ViTMatte [55] | Trimap | 2.17 | 7.44 | 17.82 | 4.07 | 4.06 | 11.27 | 2.29 |
| GVM [9] | Auxiliary-Free | 4.00 | 9.64 | 34.04 | 5.35 | 5.56 | 18.30 | 3.11 |
| MatAnyone 2 [53] | Segmentation Mask | 1.67 | 7.42 | 21.61 | 4.06 | 3.63 | 10.10 | 2.08 |
| **$\alpha$Depth (Ours)** | Depth | 2.17 | 7.24 | 20.20 | 3.89 | 4.09 | 11.88 | 2.31 |

Table 4: **Ablation study**. (a) Effects of Alpha Estimation (AE), Circular Alpha Representation (CAR), and Layered Representation (LR) on warping performance. (b) Impacts of different strategies on alpha matting. The best and second best results are marked. Please see Sec. F for visual results.

(a) $\alpha$Depth ablation.

| ID | AE | CAR | LR | Marvel-10K | | |
|---|---|---|---|---|---|---|
| | | | | S-SSIM ↑ | LPIPS ↓ | DISTS ↓ |
| A#1 | | | | 0.6718 | 0.1337 | 0.0910 |
| A#2 | ✓ | | | 0.7165 | 0.1346 | 0.0869 |
| A#3 | ✓ | ✓ | | 0.7214 | 0.1280 | 0.0819 |
| A#4 | ✓ | ✓ | ✓ | 0.7636 | 0.1190 | 0.0738 |

(b) CAR ablation.

| ID | Strategies | P3M-10K | | |
|---|---|---|---|---|
| | | SAD ↓ | Grad ↓ | Conn ↓ |
| B#1 | Vanilla Alpha Representation | 4.47 | 12.41 | 2.42 |
| B#2 | $\ell_1$ Loss Only | 4.36 | 12.54 | 2.48 |
| B#3 | Single-Stage Training | 4.29 | 11.77 | 2.43 |
| B#4 | **CAR (Ours)** | 4.09 | 11.88 | 2.31 |

### 4.3 Alpha Matting

We evaluate our Circular Alpha Representation (CAR) against trimap-based (ViTMatte [55]), mask-based (MatAnyone 2 [53]), and auxiliary-free (GVM [9]) baselines. Since CAR predicts alpha values at local soft boundaries without defining global foreground/background, we apply circular alpha encoding Eq. (2) to ground-truth labels and the predictions of all baselines for fair comparisons (please see Sec. A.2 for more details). Following ViTMatte [55], we compute standard alpha metrics (SAD, Grad, Conn) exclusively within the unknown regions of trimaps. As shown in Tab. 3 and Fig. 8, CAR performs comparably to state-of-the-art methods without requiring manual guidance. Furthermore, in contrast to prior matting methods, our design is able to handle complex multi-target scenes (Fig. 3b) and capture intra-object soft boundaries (*e.g.*, Figs. 8 top and 1 top).

### 4.4 Ablation Study

To validate our design choices, we analyze both warping and alpha matting performance in Tab. 4. **(i) $\alpha$Depth Ablation**. Compared with the baseline directly using VDA depth [3] for warping (A#1), estimating only foreground information (*i.e.*, foreground color, depth, and alpha) at soft boundaries mitigates background bleeding and improves structural recovery (S-SSIM of A#2 in Tab. 4a). Replacing the vanilla alpha representation with our proposed CAR better handles complex scenes, yielding consistent gains (A#3 *v.s.* A#2). By additionally estimating background information, our layered $\alpha$Depth representation achieves the best warping performance via soft boundary decomposition (A#4). **(ii) CAR Ablation**. Tab. 4b validates the contributions of matting loss $\mathcal{L}_{\text{m}}$ (B#2 *v.s.* B#4) and two-stage training (B#3 *v.s.* B#4) in alpha matting performance. Unlike vanilla alpha representation

Table 5: **Synergy between circular alpha and layered representation**. We follow the experimental setup in Tab. 4a and compare our circular alpha against vanilla alpha in terms of warping performance. The best and second best results are marked.

| Strategies | Marvel-10K | | |
|---|---|---|---|
| | S-SSIM ↑ | LPIPS ↓ | DISTS ↓ |
| Vanilla Alpha | 0.7165 | 0.1346 | 0.0869 |
| Vanilla Alpha + Layered Representation | 0.7207 | 0.1294 | 0.0832 |
| Circular Alpha | 0.7214 | 0.1280 | 0.0819 |
| Circular Alpha + Layered Representation | 0.7636 | 0.1190 | 0.0738 |

Table 6: **Impact of alpha maps on stereo conversion**. We conduct experiments on Marvel-10K using MatAnyone 2 [53] in place of our circular alpha estimation. Note that MatAnyone 2 requires masks as guidance, while $\alpha$Depth operates without any manual guidance. Best results are marked.

| Stage | Alpha Source | PSNR ↑ | SSIM ↑ | S-PSNR ↑ | S-SSIM ↑ | LPIPS ↓ | DISTS ↓ | FVD ↓ |
|---|---|---|---|---|---|---|---|---|
| Warping Only | MatAnyone 2 | 27.11 | 0.9055 | 24.90 | 0.6962 | 0.1305 | 0.0865 | 9.41 |
| | **$\alpha$Depth** | 27.38 | 0.9094 | 28.68 | 0.7636 | 0.1190 | 0.0738 | 7.13 |
| Stereo Conversion | MatAnyone 2 | 33.04 | 0.9312 | 27.22 | 0.8122 | 0.0730 | 0.0317 | 1.97 |
| | **$\alpha$Depth** | 33.31 | 0.9326 | 28.46 | 0.8373 | 0.0699 | 0.0297 | 1.72 |

that often suffers from unstable predictions due to alpha valley issues (Fig. 3b), our CAR focuses on local soft boundaries and delivers the best matting performance (B#1 *v.s.* B#4).

## 5 Method Analysis

This section investigates the synergistic effect of our core components and the impact of alpha representation in stereo conversion.

**Synergistic Design**. Tab. 5 shows that our circular alpha and layered representation are mutually indispensable: recovering accurate soft boundary details requires both clean layered decomposition (color and depth, provided by layered representation) and a precise blending weight (enabled by circular alpha). Adding layered representation to a vanilla alpha baseline yields only marginal gains (*e.g.*, 0.0042 S-SSIM gain in Tab. 5), whereas incorporating layered representation alongside circular alpha produces a substantially larger improvement (*e.g.*, 0.0422 S-SSIM gain). This contrast arises because the alpha valley issue degrades vanilla alpha estimation, causing blending errors that propagate through the compositing process regardless of the accuracy of the underlying layered content. Consequently, the performance benefit of layered representation only fully manifests once circular alpha resolves the alpha manifold discontinuity, making their interaction synergistic rather than additive.

**Alpha Representation**. A straightforward baseline to our automatic decomposition is providing the layered pipeline with alpha mattes from a state-of-the-art matting method. To test this, we manually annotate segmentation masks on Marvel-10K as guidance for MatAnyone 2 [53] and substitute its mattes into our conversion pipeline, evaluating both the warped intermediates and the final stereo results. As shown in Tab. 6, even with manually provided guidance, the complex cinematic scenes of Marvel-10K degrade the matting quality of MatAnyone 2, which in turn hinders accurate soft boundary synthesis in the target view. By contrast, $\alpha$Depth consistently outperforms this baseline across all metrics in both stages, despite requiring no user guidance.

## 6 Conclusion

This paper proposes $\alpha$Depth, a layered representation designed to resolve depth ambiguity and color mixing at soft boundaries for stereo conversion. By leveraging our Circular Alpha Representation (CAR), $\alpha$Depth bypasses the discontinuities of vanilla alpha in complex scenes, enabling automatic scene-level decomposition in a single forward pass. Extensive experiments verify the effectiveness of CAR, showing state-of-the-art boundary fidelity of $\alpha$Depth in stereo conversion.

## References


[1] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *ICCV*, 2025.

[2] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *ICLR*, 2025.

[3] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, pages 22831–22840, 2025.

[4] Peng Dai, Feitong Tan, Qiangeng Xu, David Futschik, Ruofei Du, Sean Fanello, Xiaojuan Qi, and Yinda Zhang. Svg: 3d stereoscopic video generation via denoising frame matrix. *arXiv preprint arXiv:2407.00367*, 2024.

[5] Yutong Dai, Brian Price, He Zhang, and Chunhua Shen. Boosting robustness of image matting with context assembling and strong data augmentation. In *CVPR*, pages 11707–11716, 2022.

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *PAMI*, 44(5):2567–2581, 2022.

[7] Mikhail Erofeev, Yury Gitman, Dmitriy S Vatolin, Alexey Fedorov, and Jue Wang. Perceptually motivated benchmark for video matting. In *BMVC*, volume 1, page 2, 2015.

[8] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. In *NeurIPS*, 2024.

[9] Yongtao Ge, Kangyang Xie, Guangkai Xu, Li Ke, Mingyu Liu, Longtao Huang, Hui Xue, Hao Chen, and Chunhua Shen. Generative video matting. In *SIGGRAPH*, pages 1–10, 2025.

[10] Michal Geyer, Omer Tov, Linyi Jin, Richard Tucker, Inbar Mosseri, Tali Dekel, and Noah Snavely. Eye2eye: A simple approach for monocular-to-stereo video synthesis. *arXiv preprint arXiv:2505.00135*, 2025.

[11] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024.

[12] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, volume 36, pages 1140–1147, 2022.

[13] Beomyoung Kim, Chanyong Shin, Joonhyun Jeong, Hyungsik Jung, Se-Yun Lee, Sewhan Chun, Dong-Hyun Hwang, and Joonsang Yu. Zim: Zero-shot image matting for anything. In *ICCV*, pages 23828–23838, 2025.

[14] Frank L Kooi and Alexander Toet. Visual comfort of binocular and 3d displays. *Displays*, 25(2-3):99–108, 2004.

[15] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *CVPR*, pages 1775–1785, 2024.

[16] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *ACMMM*, pages 3501–3509, 2021.

[17] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. Bridging composite and real: towards end-to-end deep image matting. *IJCV*, 130(2):246–266, 2022.

[18] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. In *IJCAI*, 2021.

[19] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018.

[20] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[21] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *CVPR*, pages 8762–8771, 2021.

[22] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *WACV*, pages 238–247, 2022.

[23] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024.

[24] Anna Lischke, Guofei Pang, Mamikon Gulian, Fangying Song, Christian Glusa, Xiaoning Zheng, Zhiping Mao, Wei Cai, Mark M Meerschaert, Mark Ainsworth, et al. What is the fractional laplacian? a comparative review with new results. *Journal of Computational Physics*, 404:109009, 2020.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[26] Lukas Mehl, Andrés Bruhn, Markus Gross, and Christopher Schroers. Stereo conversion with disparity-aware warping, compositing and inpainting. In *WACV*, pages 4260–4269, 2024.

[27] Nando Metzger, Prune Truong, Goutam Bhat, Konrad Schindler, and Federico Tombari. Elastic3d: Controllable stereo video conversion with guided latent decoding. In *CVPR*, 2026.

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[29] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *CVPR*, pages 5437–5446, 2020.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[31] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *CVPR*, pages 11696–11706, 2022.

[32] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.

[33] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, pages 10106–10116, 2024.

[34] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *CVPR*, 2020.

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021.

[36] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *PAMI*, 44(3):1623–1637, 2022.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[39] Guibao Shen, Yihua Du, Wenhang Ge, Jing He, Chirui Chang, Donghao Zhou, Zhen Yang, Luozhou Wang, Xin Tao, and Ying-Cong Chen. Stereopilot: Learning unified and efficient stereo conversion via generative priors. *arXiv preprint arXiv:2512.16915*, 2025.

[40] Nina Shvetsova, Goutam Bhat, Prune Truong, Hilde Kuehne, and Federico Tombari. M2svid: End-to-end inpainting and refinement for monocular-to-stereo video conversion. *arXiv preprint arXiv:2505.16565*, 2025.

[41] Filippo Speranza, Wa James Tam, Ron Renaud, and Namho Hur. Effect of disparity and motion on visual comfort of stereoscopic images. In *Stereoscopic Displays and Virtual Reality Systems XIII*, volume 6055, pages 94–103. SPIE, 2006.

[42] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[43] Lezhong Wang, Jeppe Revall Frisvad, Mark Bo Jensen, and Siavash Arjomand Bigdeli. Stereodiffusion: Training-free stereo image generation using latent diffusion models. In *CVPR*, pages 7416–7425, 2024.

[44] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *CVPR*, pages 5261–5271, 2025.

[45] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In *NeurIPS*, 2025.

[46] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *CVPR*, pages 5249–5260, 2025.

[47] Hongyu Wen and Jia Deng. Seegroup: Multi-layer depth estimation of transparent surfaces via self-determined grouping. *arXiv preprint arXiv:2605.28735*, 2026.

[48] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *ECCV*, pages 842–857. Springer, 2016.

[49] Gangwei Xu, Haotong Lin, Hongcheng Luo, Xianqi Wang, Jingfeng Yao, Lianghui Zhu, Yuechuan Pu, Cheng Chi, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Sida Peng, and Xin Yang. Pixel-perfect depth with semantics-prompted diffusion transformers. In *NeurIPS*, 2025.

[50] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, pages 2970–2979, 2017.

[51] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pages 10371–10381, 2024.

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024.

[53] Peiqing Yang, Shangchen Zhou, Kai Hao, and Qingyi Tao. Matanyone 2: Scaling video matting via a learned quality evaluator. In *CVPR*, 2026.

[54] Peiqing Yang, Shangchen Zhou, Jixin Zhao, Qingyi Tao, and Chen Change Loy. Matanyone: Stable video matting with consistent memory propagation. In *CVPR*, pages 7299–7308, 2025.

[55] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103:102091, 2024.

[56] Wei Yin, Xinlong Wang, Chunhua Shen, Yifan Liu, Zhi Tian, Songcen Xu, Changming Sun, and Dou Renyin. Diversedepth: Affine-invariant depth prediction using diverse data. *arXiv preprint arXiv:2002.00569*, 2020.

[57] Songsong Yu, Yuxin Chen, Zhongang Qi, Zeke Xie, Yifan Wang, Lijun Wang, Ying Shan, and Huchuan Lu. Mono2stereo: A benchmark and empirical study for stereo conversion. In *CVPR*, pages 21847–21856, 2025.

[58] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[60] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. In *NeurIPS*, 2024.

[61] Xiang Zhang, Yang Zhang, Lukas Mehl, Markus Gross, and Christopher Schroers. High-fidelity novel view synthesis via splatting-guided diffusion. In *SIGGRAPH*, SIGGRAPH Conference Papers '25, New York, NY, USA, 2025. Association for Computing Machinery.

[62] Xiang Zhang, Yang Zhang, Lukas Mehl, Markus Gross, and Christopher Schroers. Guardians of the hair: Rescuing soft boundaries in depth, stereo, and novel views. In *CVPR*, 2026.

[63] Sijie Zhao, Wenbo Hu, Xiaodong Cun, Yong Zhang, Xiaoyu Li, Zhe Kong, Xiangjun Gao, Muyao Niu, and Ying Shan. Stereocrafter: Diffusion-based generation of long and high-fidelity stereoscopic 3d from monocular videos. *arXiv preprint arXiv:2409.07447*, 2024.

[64] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), July 2018.

# Appendix

We provide more technical details, experimental results, ablation studies, and qualitative visualizations to support the contributions of our $\alpha$Depth approach. Detailed contents are listed as follows:

# Contents


**A More Implementation Details** **15**

A.1 Layered Warping with $\alpha$Depth Representation . . . . . 15

A.2 Matting Evaluation Details . . . . . 16

A.3 Training Data Curation . . . . . 16

A.4 Matting Loss $\mathcal{L}_{\mathrm{m}}$ . . . . . 16

**B More Experimental Results** **17**

B.1 Pixel-Level Metrics on Full Image . . . . . 17

B.2 Temporal Stability and Stereo Comfort . . . . . 17

B.3 Vanilla Alpha Representation *v.s.* Circular Alpha Representation . . . . . 19

B.4 Matting Performance under Different Depth Inputs . . . . . 20

B.5 Performance under Different Camera Trajectories . . . . . 20

B.6 Computational Complexity . . . . . 21

**C More Ablation Studies** **21**

C.1 Impact of Multi-Branch Decoder . . . . . 21

C.2 Impact of Semantic Encoder . . . . . 22

C.3 Impact of Depth Edge Extraction . . . . . 23

**D Limitations and Discussions** **23**

**E Discussion of Societal Impacts** **24**

**F Visualization of Ablation Models** **24**

**G Visualization of $\alpha$Depth Results** **24**

**H More Visual Comparisons** **25**

H.1 Stereo Conversion . . . . . 25

H.2 Alpha Matting . . . . . 25


# A More Implementation Details

## A.1 Layered Warping with $\alpha$Depth Representation

Previous stereo conversion approaches often employ techniques like softmax splatting [29] for view transformation [26, 40, 57, 62]. To support the layered representation of $\alpha$Depth, we extend the softmax splatting to first jointly project the foreground layer with premultiplied alpha, *i.e.*,

$$\tilde{I}_{\alpha\mathrm{FG}}, \tilde{\alpha} = \mathrm{Project}(\{\hat{\alpha} \odot \hat{I}_{\mathrm{FG}}, \hat{\alpha}\}, \hat{D}_{\mathrm{FG}}), \tag{6}$$

where $\mathrm{Project}(\cdot)$ represents depth-guided softmax splatting [26] to handle occlusions. The joint projection ensures that $\tilde{\alpha}$ aligns with the foreground color $\tilde{I}_{\mathrm{FG}}$. We then separately project the background layer

$$\tilde{I}_{\mathrm{BG}} = \mathrm{Project}(\hat{I}_{\mathrm{BG}}, \hat{D}_{\mathrm{BG}}). \tag{7}$$

Finally, we generate the warped view $\tilde{I}$ via alpha compositing on soft boundary regions:

$$\tilde{I} = \tilde{I}_{\alpha\mathrm{FG}} + (1 - \tilde{\alpha})\tilde{I}_{\mathrm{BG}}. \tag{8}$$

Since our $\alpha$Depth model only estimates layered information for soft boundary regions, with zero $\tilde{\alpha}$ for opaque regions, Eq. (8) performs alpha composition only on soft boundaries. Thus, the warped image $\tilde{I}$ preserves the geometry estimated from state-of-the-art depth models [49, 45, 20], while recovering high-fidelity structures on soft boundaries.

### A.2 Matting Evaluation Details

We provide more details for the matting evaluation protocol used in Sec. 4.3. Since our $\alpha$Depth representation uses zero to represent opaque regions without differentiating foreground or background, directly computing alpha metrics cannot reflect our detail extraction performance on soft boundary regions. Thus, we first apply circular alpha encoding (Eq. (2)) to project ground-truth alpha labels and the alpha estimation results of all baselines into the continuous trigonometric space. This maps the foreground ($\alpha = 1$) and background ($\alpha = 0$) to the same coordinate $(0, 1)$ in the $(\alpha_{\sin}, \alpha_{\cos})$ space, benefiting the evaluation on soft boundary regions. Then, we compute alpha metrics on the trigonometric space, *i.e.*,

$$\begin{aligned} \mathrm{SAD} &= 0.5 \cdot \mathrm{SAD}(\hat{\alpha}_{\sin}, \alpha_{\sin}) + 0.5 \cdot \mathrm{SAD}(\hat{\alpha}_{\cos}, \alpha_{\cos}), \\ \mathrm{Grad} &= 0.5 \cdot \mathrm{Grad}(\hat{\alpha}_{\sin}, \alpha_{\sin}) + 0.5 \cdot \mathrm{Grad}(\hat{\alpha}_{\cos}, \alpha_{\cos}), \\ \mathrm{Conn} &= 0.5 \cdot \mathrm{Conn}(\hat{\alpha}_{\sin}, \alpha_{\sin}) + 0.5 \cdot \mathrm{Conn}(\hat{\alpha}_{\cos}, \alpha_{\cos}), \end{aligned} \tag{9}$$

where $(\hat{\alpha}_{\sin}, \hat{\alpha}_{\cos})$ and $(\alpha_{\sin}, \alpha_{\cos})$ represent the estimated result and ground-truth label, respectively. $\mathrm{SAD}(\cdot), \mathrm{Grad}(\cdot), \mathrm{Conn}(\cdot)$ indicate the commonly used Sum of Absolute Differences (SAD), Gradient loss (Grad), and Connectivity loss (Conn) [55]. We also follow ViTMatte to only compute the metrics on the unknown regions of the official trimap in the matting datasets [55].

For the input guidance to matting baselines, we employ the official trimaps for the trimap-based method ViTMatte [55]. The recent approach MatAnyone 2 requires binary segmentation masks as guidance, which are not available in AIM-500 [18] and P3M-10K [16]. Thus, we generate binary masks by thresholding the ground-truth alpha maps, *i.e.*, $\mathbb{I}(0.5 \leq \alpha)$. Compared with previous methods, our $\alpha$Depth model directly infers soft boundary regions from image semantics and geometry layout, achieving comparable matting performance (see Tab. 3) without requiring manual guidance.

### A.3 Training Data Curation

We provide more implementation details in training data curation (Fig. 6). Given the original foreground image $I_{\mathrm{FG}}^{\mathrm{ori}}$ (unpremultiplied, *i.e.*, not mixed with background colors) from matting datasets and the original background image $I_{\mathrm{BG}}^{\mathrm{ori}}$ from image datasets, we first apply alpha composition to generate the input image,

$$I_{\mathrm{IN}} = \alpha I_{\mathrm{FG}}^{\mathrm{ori}} + (1 - \alpha) I_{\mathrm{BG}}^{\mathrm{ori}}. \tag{10}$$

For the ground-truth layered color $I_{\mathrm{FG}}, I_{\mathrm{BG}}$, we perform masked blending based on the binary foreground and background masks $M_{\mathrm{FG}} = \mathbb{I}(\alpha \geq \alpha_{\mathrm{th}}), M_{\mathrm{BG}} = \mathbb{I}(\alpha \geq 1 - \alpha_{\mathrm{th}})$, *i.e.*,

$$\begin{aligned} I_{\mathrm{FG}} &= M_{\mathrm{FG}} \odot I_{\mathrm{FG}}^{\mathrm{ori}} + (1 - M_{\mathrm{FG}}) \odot I_{\mathrm{BG}}^{\mathrm{ori}}, \\ I_{\mathrm{BG}} &= M_{\mathrm{BG}} \odot I_{\mathrm{FG}}^{\mathrm{ori}} + (1 - M_{\mathrm{BG}}) \odot I_{\mathrm{BG}}^{\mathrm{ori}}. \end{aligned} \tag{11}$$

Regarding depth data generation, we first follow HairGuard to obtain high-quality depth $D_{\mathrm{FG}}^{\mathrm{ori}}, D_{\mathrm{BG}}^{\mathrm{ori}}$ using pre-trained depth models, and then generate the input depth map $D_{\mathrm{IN}}$ via depth composition [62]. For the ground-truth depth labels $D_{\mathrm{FG}}, D_{\mathrm{BG}}$, we perform masked blending based on the same foreground and background masks $M_{\mathrm{FG}}, M_{\mathrm{BG}}$, *i.e.*,

$$\begin{aligned} D_{\mathrm{FG}} &= M_{\mathrm{FG}} \odot D_{\mathrm{FG}}^{\mathrm{ori}} + (1 - M_{\mathrm{FG}}) \odot D_{\mathrm{BG}}^{\mathrm{ori}}, \\ D_{\mathrm{BG}} &= M_{\mathrm{BG}} \odot D_{\mathrm{FG}}^{\mathrm{ori}} + (1 - M_{\mathrm{BG}}) \odot D_{\mathrm{BG}}^{\mathrm{ori}}. \end{aligned} \tag{12}$$

As shown in Fig. 6, the generated $I_{\mathrm{FG}}/D_{\mathrm{FG}}$ and $I_{\mathrm{BG}}/D_{\mathrm{BG}}$ preserve foreground and background information on soft boundary regions, respectively.

### A.4 Matting Loss $\mathcal{L}_{\mathrm{m}}$

To enhance the detail extraction performance of $\alpha$Depth, we adopt the matting loss $\mathcal{L}_{\mathrm{m}}$ introduced in ViTMatte [55]. Specifically, $\mathcal{L}_{\mathrm{m}}$ is composed of three terms:

$$\mathcal{L}_{\mathrm{m}} = \mathcal{L}_1 + \mathcal{L}_{lap} + \mathcal{L}_{gp}, \tag{13}$$

where $\mathcal{L}_{lap}$ and $\mathcal{L}_{gp}$ denote the Laplacian loss [24] and the gradient loss [5], respectively. Since this matting loss is designed for single-channel data (*e.g.*, alpha mattes), it can be directly applied to supervise our depth predictions. For the foreground and background color outputs, we compute $\mathcal{L}_{\mathrm{m}}$ independently across each color channel and average the results.

Table 7: **Pixel-level metrics on full image**. The best and second best results are marked.

| Method | Mono2Stereo [57] | | Marvel-10K [62] | |
|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| **Stereo Conversion** | | | | |
| StereoDiffusion [43] | 26.34 | 0.7541 | 26.04 | 0.7656 |
| Mono2Stereo [57] | 28.15 | 0.7980 | 28.19 | 0.8132 |
| StereoCrafter [63] | 28.06 | 0.7855 | 29.27 | 0.8340 |
| SplatDiff [61] | 30.40 | 0.8909 | 32.37 | 0.9284 |
| HairGuard [62] | 30.88 | 0.8911 | 33.09 | 0.9316 |
| **$\alpha$Depth (Ours)** | 31.03 | 0.8942 | 33.31 | 0.9326 |
| **Warping Performance** | | | | |
| DAv2 [52] / VDA [3] | 29.35 | 0.8888 | 27.04 | 0.9055 |
| DAv2 / VDA + HairGuard [62] | 29.35 | 0.8889 | 27.06 | 0.9068 |
| **DAv2 / VDA + $\alpha$Depth (Ours)** | 29.38 | 0.8891 | 27.38 | 0.9094 |

Table 8: **Statistical confidence analysis of full-image metrics on stereo conversion.** $\bar{\Delta}$ represents the mean difference with its 95% paired-bootstrap Confidence Interval (CI). We also report the one-sided Wilcoxon signed-rank p-value, and the fraction of samples where our method is better, *i.e.*, the win rate (Win%).

| Dataset | Metric | $\bar{\Delta}$ (95% CI) | $p$ (Wilcoxon) | Win% |
|---|---|---|---|---|
| Marvel-10K | PSNR↑ | $+0.2280$ $[+0.1360, +0.3220]$ | $9.0 \times 10^{-10}$ | 62% |
| | SSIM↑ | $+0.0010$ $[+0.0006, +0.0013]$ | $8.7 \times 10^{-9}$ | 62% |
| | LPIPS↓ | $-0.0035$ $[-0.0040, -0.0031]$ | $6.3 \times 10^{-43}$ | 78% |
| | DISTS↓ | $-0.0025$ $[-0.0028, -0.0021]$ | $5.9 \times 10^{-39}$ | 78% |
| Mono2Stereo | PSNR↑ | $+0.1420$ $[+0.1210, +0.1620]$ | $3.6 \times 10^{-129}$ | 76% |
| | SSIM↑ | $+0.0031$ $[+0.0029, +0.0032]$ | $4.9 \times 10^{-281}$ | 86% |
| | LPIPS↓ | $-0.0031$ $[-0.0033, -0.0029]$ | $9.2 \times 10^{-191}$ | 76% |
| | DISTS↓ | $-0.0027$ $[-0.0029, -0.0025]$ | $1.5 \times 10^{-186}$ | 75% |

# B More Experimental Results

## B.1 Pixel-Level Metrics on Full Image

Following the experimental settings in Tabs. 1 and 2, we additionally provide pixel-level metrics computed on the full image. Tab. 7 verifies the state-of-the-art performance of $\alpha$Depth in both warping and stereo conversion. To assess the reliability of the reported improvements, we further perform a paired statistical analysis against HairGuard on stereo conversion. Since all methods are evaluated on identical inputs, we form per-sample paired differences $\Delta_i = \mathrm{Metric}_i(\text{Ours}) - \mathrm{Metric}_i(\text{HairGuard})$ for each metric and report three complementary statistics: (i) a 95% confidence interval on the mean difference $\bar{\Delta}$, obtained by the percentile paired bootstrap with 10,000 resamples drawn over samples with replacement; (ii) a one-sided Wilcoxon signed-rank test, which is distribution-free and robust to the heavy-tailed difference distributions; and (iii) the win rate, *i.e.*, the fraction of samples on which our method outperforms HairGuard. All analyses use a fixed random seed, and the direction of the one-sided tests follows each metric's polarity (higher is better for PSNR/SSIM, lower for LPIPS/DISTS). As shown in Tab. 8, all 95% confidence intervals exclude zero, all Wilcoxon p-values are below $10^{-8}$, and our method wins on 62–86% of individual samples, confirming that the improvements are statistically reliable and consistent.

## B.2 Temporal Stability and Stereo Comfort

Beyond per-frame quality, temporal stability and stereo viewing comfort are crucial for high-quality stereo conversion. Let $\hat{I}_t$ and $I_t$ denote the synthesized view and the ground-truth view of frame $t$, respectively. We first assess temporal stability using video-oriented metrics:

- **Warping Error ($E_{\text{warp}}$)**. Temporal flicker is evaluated via the warping error, which measures photometric agreement between consecutive predicted frames after motion compensation, *i.e.*,

$$E_{\text{warp}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{p} \in \mathcal{V}} \left\| \hat{I}_{t+1}(\mathbf{p}) - \overrightarrow{\hat{I}_t}(\mathbf{p}) \right\|_1, \tag{14}$$

Table 9: **Comparison of temporal stability and stereo comfort metrics** on the Marvel-10K dataset. The best and second best results are marked.

| **Method** | $E_{\text{warp}} \downarrow$ | RFlick-B $\downarrow$ | $E_{\text{disp}} \downarrow$ | TDV $\downarrow$ | VDisp $\downarrow$ |
|---|---|---|---|---|---|
| StereoCrafter | 0.681 | 0.547 | 18.56 | 5.309 | 0.285 |
| SplatDiff | 0.713 | 0.466 | 7.29 | 2.242 | 0.078 |
| HairGuard | 0.589 | 0.464 | 6.87 | 2.044 | 0.077 |
| $\alpha$**Depth (Ours)** | 0.588 | 0.445 | 6.58 | 2.029 | 0.075 |

where $\overrightarrow{\hat{I}_t}$ denotes frame $t$ warped to frame $t{+}1$ with optical flow estimated on the reference sequence, $\mathbf{p} = (u, v)^\top$ indicates pixel coordinates, and $\mathcal{V}$ excludes occluded pixels.

- **Boundary Residual Flicker (RFlick-B)**. Boundary regions of synthesized videos are prone to temporal flicker, and thus we define a boundary residual flicker metric RFlick-B for boundary stability evaluation. Let $B_t$ denote the reference disparity edge region of frame $t$, and let

$$r_t \;=\; \frac{1}{|B_t|} \sum_{\mathbf{p} \in B_t} \left\| \hat{I}_t(\mathbf{p}) - I_t(\mathbf{p}) \right\|_1 \tag{15}$$

be the mean boundary residual at frame $t$. RFlick-B is then the temporal standard deviation of these residuals,

$$\text{RFlick-B} \;=\; \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(r_t - \bar{r}\right)^2}, \quad \bar{r} \;=\; \frac{1}{T} \sum_{t=1}^{T} r_t, \tag{16}$$

where $T$ is the number of frames. The metric focuses on the temporal instability of boundary errors, which often introduces perceptible visual artifacts, such as flickering hair.

Since image-level metrics are largely insensitive to disparity errors (*e.g.*, a view with a plausible appearance but wrong parallax scores well on PSNR), we employ stereo metrics that extract disparity from the synthesized pair with an off-the-shelf stereo matcher [46] and assess it directly. In the following, $\hat{\delta}_t$ and $\delta_t$ denote the disparities estimated from the generated and reference stereo pairs of frame $t$, respectively.

- **Disparity Error ($E_{\text{disp}}$)**. We follow Elastic3D [27] to measure the geometric accuracy of the induced parallax as the mean absolute deviation between the two, *i.e.*,

$$E_{\text{disp}} \;=\; \frac{1}{THW} \sum_{t=1}^{T} \sum_{v=1}^{H} \sum_{u=1}^{W} \left| \hat{\delta}_t(\mathbf{p}) - \delta_t(\mathbf{p}) \right|, \tag{17}$$

directly quantifying whether the synthesized view places content at the intended depth.

- **Temporal Disparity Variation (TDV)**. Temporal fluctuations of the overall depth placement are known to induce visual discomfort [41], and thus we quantify them as the temporal disparity variation TDV, *i.e.*,

$$\text{TDV} \;=\; \frac{1}{T-1} \sum_{t=2}^{T} \left| \bar{\delta}_t - \bar{\delta}_{t-1} \right|, \tag{18}$$

where $\bar{\delta}_t$ denotes the spatial mean of $\hat{\delta}_t$.

- **Vertical Disparity (VDisp)**. A rectified stereo pair must contain purely horizontal parallax, and any vertical component violates epipolar geometry and is a common source of eye strain [14]. We measure it as the mean vertical disparity VDisp:

$$\text{VDisp} \;=\; \frac{1}{THW} \sum_{t=1}^{T} \sum_{v=1}^{H} \sum_{u=1}^{W} \left| F_t^y(\mathbf{p}) \right|, \tag{19}$$

where $F_t^y$ denotes the vertical component of the optical flow estimated between the left and right views of frame $t$.

As presented in Tab. 9, $\alpha$Depth achieves superior performance across all five metrics. The margin over HairGuard is particularly pronounced on RFlick-B and $E_{\text{disp}}$, demonstrating that our layered representation effectively stabilizes high-frequency boundary content and maps it to more plausible disparities.

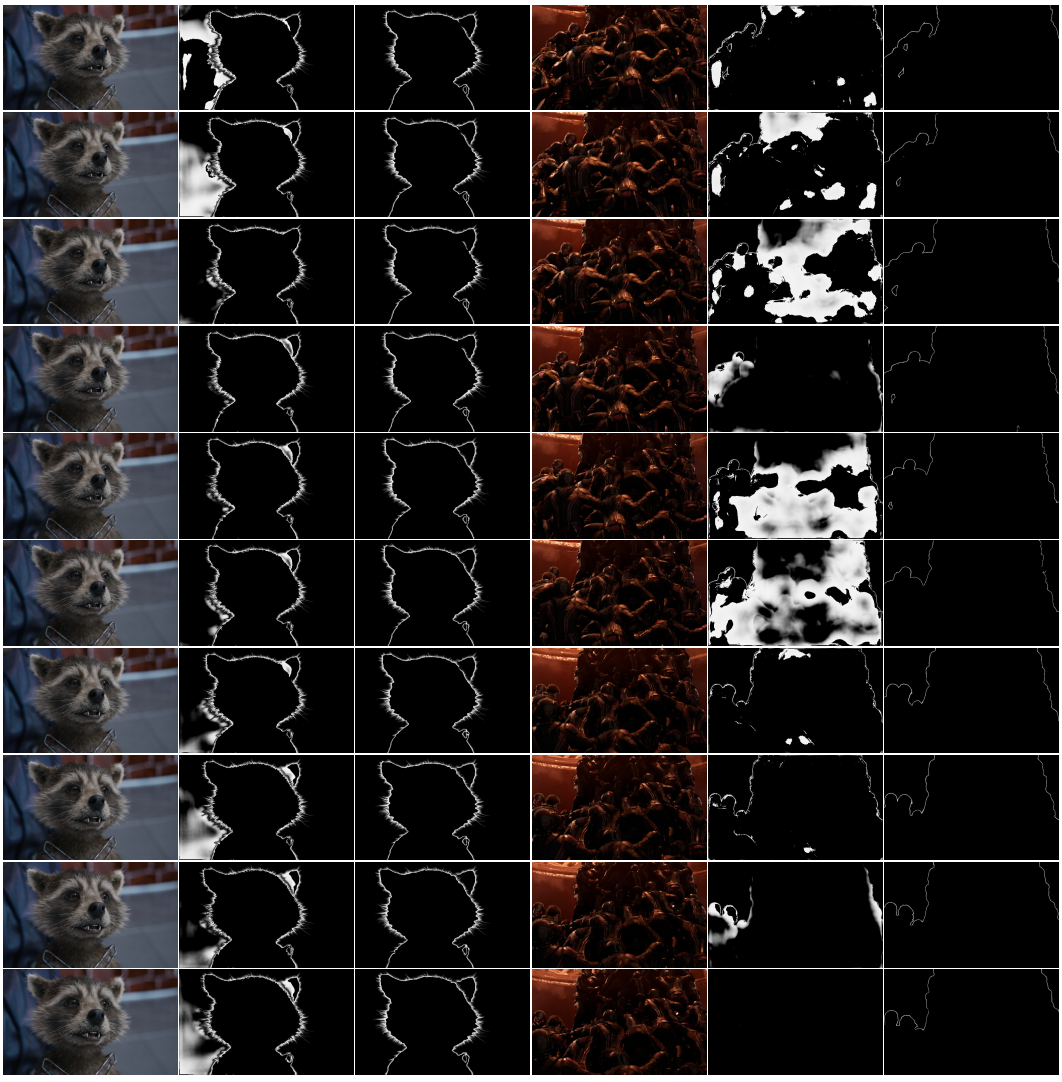



Figure 9: **Stability comparisons** between vanilla alpha representation and circular alpha representation (CAR). Vanilla alpha representation often suffers from alpha valley issues and thus produces unstable results. By contrast, our CAR shows consistent performance when processing video inputs. Regions outside $\alpha \in [0.02, 0.98]$ are masked out for better comparison.

### B.3 Vanilla Alpha Representation *v.s.* Circular Alpha Representation

Due to the discontinuity in alpha labels (*e.g.*, see the left part of Fig. 5), learning to predict vanilla alpha representation in complex scenes often suffers from alpha valley issues (Fig. 3b), resulting in inaccurate alpha estimation. To verify this, we employ the $\alpha$Depth network and train a variant model for vanilla alpha prediction under the same training settings. Fig. 9 compares the performance of vanilla alpha representation and our Circular Alpha Representation (CAR) under video inputs. It is evident that vanilla alpha representation often struggles at the intersections of multiple foreground targets, leading to unstable performance. In contrast, by focusing on local soft boundaries, our CAR achieves remarkable temporal consistency despite relying on an image-based model. We further conduct quantitative evaluation on two widely used video matting benchmarks, VideoMatte [21] and YouTubeMatte [54], and additionally report the temporal consistency metric dtSSD [7]. As shown in Tab. 10, CAR consistently outperforms vanilla alpha representation across all boundary precision and temporal consistency metrics on both benchmarks. In particular, it reduces dtSSD from 7.40 to 6.23 on VideoMatte and from 8.76 to 6.69 on YouTubeMatte. These results demonstrate that resolving the manifold discontinuity directly translates into significantly reduced temporal flicker and higher spatio-temporal boundary stability.

Table 10: **Ablation on the circular alpha representation for video matting** on the VideoMatte [21] and YouTubeMatte [54] datasets. Best results are marked.

| Method | VideoMatte | | | | YouTubeMatte | | | |
|---|---|---|---|---|---|---|---|---|
| | SAD ↓ | Grad ↓ | Conn ↓ | dtSSD ↓ | SAD ↓ | Grad ↓ | Conn ↓ | dtSSD ↓ |
| Vanilla Alpha | 3.07 | 3.24 | 0.73 | 7.40 | 2.01 | 4.67 | 0.95 | 8.76 |
| **Circular Alpha (Ours)** | 2.42 | 1.94 | 0.54 | 6.23 | 1.02 | 1.96 | 0.55 | 6.69 |

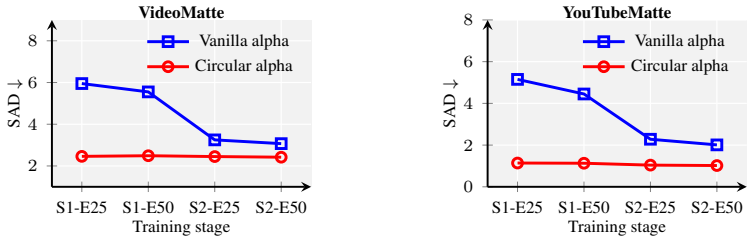



Figure 10: **Training convergence comparison** of vanilla alpha and circular alpha, measured by SAD every 25 epochs at each stage under the same training setup. SX-EY indicates training stage X and epoch Y.

The necessity of CAR is further highlighted by its training convergence. In Fig. 10, we track SAD performance on both benchmarks every 25 epochs for each training stage (SX-EY means training stage X, epoch Y), under the same training setup. By projecting the linear alpha scale onto a continuous circular manifold, CAR eliminates the non-smooth gradient landscape inherent to vanilla representations. As a result, CAR achieves accelerated convergence early in training, reaching a low SAD score of 2.46 on VideoMatte at Stage 1-Epoch 25 compared to 5.95 for vanilla alpha. Furthermore, while vanilla alpha struggles to escape local minima during early stages, CAR maintains smooth gradient flow throughout, demonstrating that resolving the alpha valley issue is critical for stable and efficient training.

### B.4 Matting Performance under Different Depth Inputs

Depth estimation performance is essential for high-quality stereo conversion because it directly influences geometry and parallax of the synthesized views. Since $\alpha$Depth focuses on soft boundary decomposition while preserving the original geometry and texture of opaque regions, it can integrate seamlessly with state-of-the-art depth estimators in a plug-and-play manner. To evaluate the robustness of $\alpha$Depth against various depth inputs, we conduct experiments on both image and video matting benchmarks, and our multimodal design remains robust to depth errors across spatial and temporal dimensions. **(i) Spatial Robustness**: As shown in Tab. 11, $\alpha$Depth achieves consistent performance across four different depth backbones. Fig. 11 (top rows) highlights three characteristic failure modes: over-smoothed discontinuities from Depth Anything V2 (DAv2), missing fine structures from Depth Pro (DPro), and incorrect depth ordering from Pixel-Perfect Depth (PPD). In all cases, $\alpha$Depth yields consistent decompositions and captures soft boundary details even where the depth inputs fail (*e.g.*, the fine dandelions in Fig. 11). This is because our dual-path encoder utilizes image priors to compensate for unreliable depth cues. **(ii) Temporal Robustness**: Tab. 12 compares per-frame depth from DAv2 against temporally consistent depth from VDA on two video matting datasets VideoMatte [21] and YouTubeMatte [54], and the performance using per-frame depth matches that of video-consistent depth across all metrics. This confirms that temporal depth flicker does not propagate into our predictions, as $\alpha$Depth anchors its estimation on the temporally stable visual sequence.

### B.5 Performance under Different Camera Trajectories

We test the performance of our $\alpha$Depth under different camera trajectories with larger viewpoint changes. Two categories of camera motions are employed: (i) We first apply a horizontal swing motion to simulate different baseline lengths in stereo conversion settings. (ii) For more flexible camera motion, we follow ReCamMaster [1] to test the robustness of our $\alpha$Depth method with 10 types of different trajectories. Tab. 13 verifies the state-of-the-art performance of $\alpha$Depth under different camera motions. Existing depth models often struggle at soft boundaries due to depth ambiguity, leading to broken structures at soft boundary regions. Although the recent approach HairGuard [62] refines depth to capture soft boundary details, the improved warping still suffers from background bleeding artifacts, as illustrated in Fig. 12. By decomposing soft boundaries via $\alpha$Depth, our method achieves the best performance in soft boundary preservation.

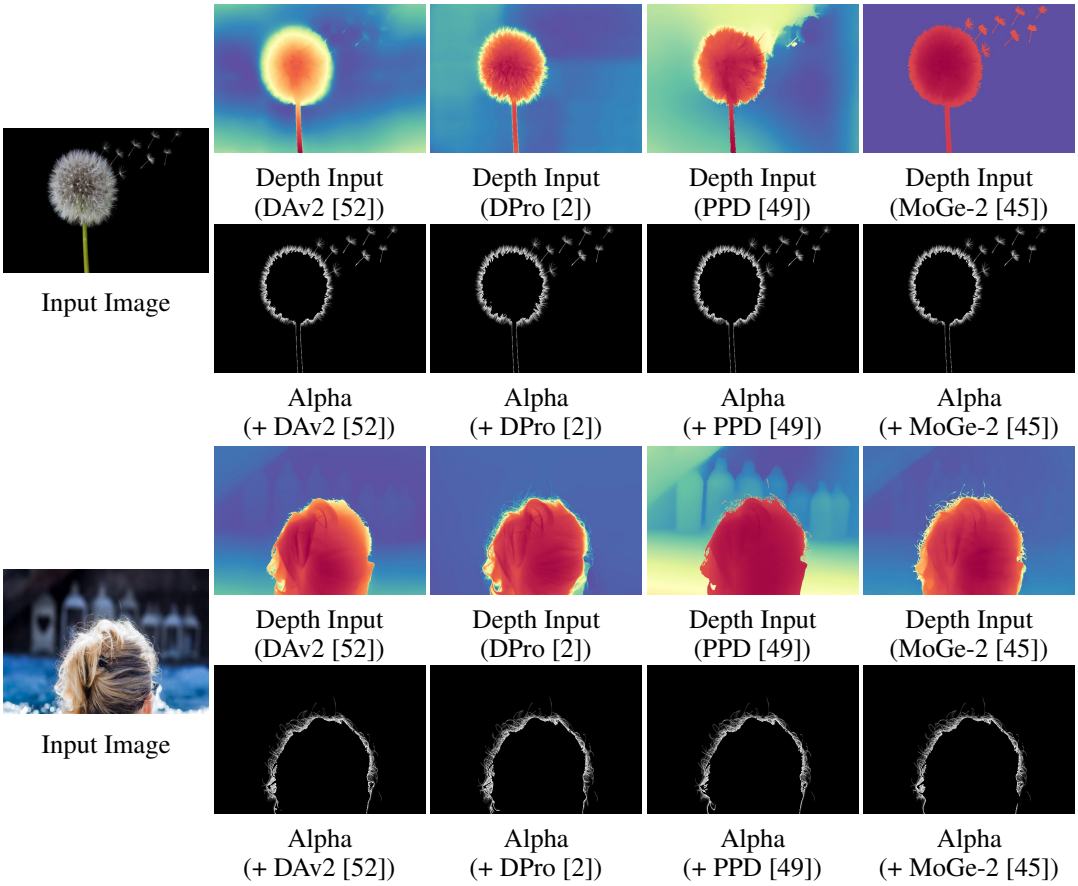



Figure 11: **Alpha estimation performance** under different depth inputs. We generate input depth using state-of-the-art models, including Depth Anything V2 (DAv2) [52], Depth Pro (DPro) [2], Pixel-Perfect Depth (PPD) [49], and MoGe-2 [45]. Despite different characteristics exhibited in depth inputs, our $\alpha$Depth shows stable performance in alpha estimation and soft boundary detail extraction.

Table 11: **Matting performance of $\alpha$Depth** with different depth models. The best and second best results are marked.

| Depth Model | AIM-500 | | | P3M-10K | | |
|---|---|---|---|---|---|---|
| | SAD ↓ | Grad ↓ | Conn ↓ | SAD ↓ | Grad ↓ | Conn ↓ |
| Depth Anything V2 [52] | 7.24 | 20.20 | 3.89 | 4.09 | 11.88 | 2.31 |
| Depth Pro [2] | 7.22 | 20.35 | 3.88 | 4.15 | 12.20 | 2.34 |
| MoGe-2 [45] | 7.34 | 21.09 | 3.95 | 4.19 | 12.47 | 2.36 |
| Pixel-Perfect Depth [49] | 7.37 | 20.76 | 3.95 | 4.18 | 12.24 | 2.36 |

### B.6 Computational Complexity

Tab. 14 breaks down the number of parameters for each component of our $\alpha$Depth network. Additionally, for an input size of $448 \times 640$, the model requires 1.94 GB of peak GPU memory and achieves an inference speed of 0.0153 seconds per image on an NVIDIA GeForce RTX 4090 GPU.

## C More Ablation Studies

### C.1 Impact of Multi-Branch Decoder

As shown in Fig. 4, we employ a multi-branch decoder to estimate different modalities, *i.e.*, alpha, depth, and color, in our $\alpha$Depth representation. This benefits the network by explicitly decoupling the distinct structural

Table 12: **Robustness to temporally inconsistent depth** on the VideoMatte [21] and YouTube-Matte [54] datasets. We compare video-consistent depth (VDA [3]) against per-frame depth (DAv2 [52]) as input to $\alpha$Depth. The temporal alpha metric dtSSD [7] is additionally adopted for evaluation, and best results are marked.

| Depth Input | VideoMatte | | | | YouTubeMatte | | | |
|---|---|---|---|---|---|---|---|---|
| | SAD ↓ | Grad ↓ | Conn ↓ | dtSSD ↓ | SAD ↓ | Grad ↓ | Conn ↓ | dtSSD ↓ |
| Video Depth (VDA) | 2.42 | 1.94 | 0.54 | 6.23 | 1.02 | 1.96 | 0.55 | 6.69 |
| Image Depth (DAv2) | 2.41 | 1.98 | 0.51 | 6.09 | 1.01 | 1.99 | 0.54 | 6.61 |

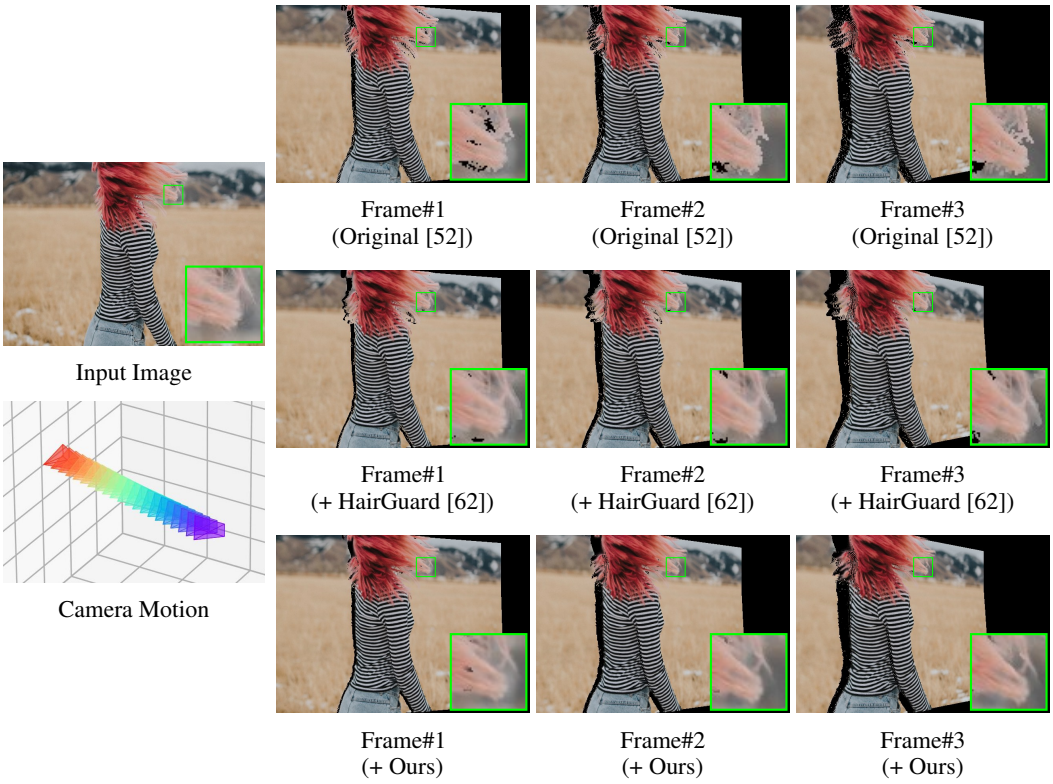



Figure 12: **Warping performance under large viewpoint changes**. This example employs the camera motion (arc left with rotation) from ReCamMaster [1]. Due to depth ambiguity in soft boundary regions, the warping results using the original depth from Depth Anything V2 [52] often contain broken structures. Although HairGuard refines depth to better preserve soft boundary details [62], its results often suffer from background bleeding. The proposed $\alpha$Depth achieves the best warping performance with high-fidelity soft boundary details.

and textural characteristics of each modality, preventing feature interference during task-specific predictions. To verify this, we train an additional variant to directly estimate $\alpha$Depth representation via a unified decoder. As demonstrated in Tab. 15, removing the multi-branch decoder leads to a consistent drop in alpha matting performance across all metrics. Specifically, the full model with the multi-branch decoder improves the SAD metric from 8.22 to 7.24 on the AIM-500 dataset [18], and from 4.36 to 4.09 on the P3M-10K dataset [16], validating its effectiveness in accurately extracting soft boundary details.

### C.2 Impact of Semantic Encoder

In the proposed $\alpha$Depth network, we employ a semantic encoder to extract high-level semantics and a detail encoder to capture soft boundary details (Fig. 4). The semantic encoder leverages pre-trained image priors to extract deep semantic features, which are essential for high-level soft boundary reasoning. Tab. 16 presents the ablation study of the semantic encoder on the Marvel-10K dataset [62], following the same experimental setups in Tab. 4a. Removing the semantic encoder results in a noticeable performance drop in warping performance, with S-PSNR decreasing from 28.68 to 27.35 and S-SSIM dropping from 0.7636 to 0.7147. This degradation

Table 13: **Warping performance (FID ↓)** under different camera trajectories on natural image matting datasets. We apply horizontal swing motion to simulate different baseline lengths in stereo conversion. We also employ 10 different camera trajectories from the evaluation protocol of ReCamMaster [1] to compare the warping performance under larger and more flexible viewpoint changes. The best and second best results are marked.

| Method | AIM-500 | | P3M-10K | |
|---|---|---|---|---|
| | Horizontal Swing | ReCamMaster Motion | Horizontal Swing | ReCamMaster Motion |
| DAv2 [52] | 41.75 | 36.99 | 49.50 | 48.47 |
| DAv2 + HairGuard [62] | 38.49 | 36.00 | 45.71 | 47.17 |
| **DAv2 + $\alpha$Depth (Ours)** | 34.43 | 34.26 | 35.47 | 43.58 |

Table 14: **Number of parameters for each component in $\alpha$Depth network.**

| Dual-Path Encoder | | Multi-Branch Decoder | | | Total |
|---|---|---|---|---|---|
| Detail Encoder | Semantic Encoder | Alpha Decoder | Color Decoder | Depth Decoder | |
| 1.17 M | 319.46 M | 2.65 M | 2.65 M | 2.65 M | 328.58 M |

indicates that high-level contextual cues and semantic understanding are vital for the network to resolve depth ambiguity and correctly decompose soft boundaries in complex scenes.

### C.3 Impact of Depth Edge Extraction

Depth edge extraction provides explicit geometric priors by extracting depth gradients from the input depth map, which serve as strong cues for soft boundary localization (Fig. 4). We empirically found that depth edges are crucial for the convergence of the network during training. Removing this component can lead to training instability or even divergence. This is likely because depth edges provide a strong initialization for soft boundary localization and thus alleviate the difficulty of multi-task learning. As shown in Tab. 16, omitting the edge extraction module severely impacts the warping performance, leading to a 2.54 dB drop in S-PSNR. This emphasizes the necessity of explicit boundary cues in guiding the network to focus on soft boundary regions.

## D Limitations and Discussions

Although $\alpha$Depth effectively resolves soft boundaries in stereo conversion, some limitations remain:

**Dependence on Initial Depth Maps**. Our method adopts a plug-and-play design that can be integrated with various state-of-the-art monocular depth estimation models. Although $\alpha$Depth effectively decomposes local foreground and background at soft boundary regions, the global scene structure and scale remain heavily dependent on the quality of the initial input depth map. If the underlying depth model fails in extreme scenarios (*e.g.*, severe geometric distortions), our approach may struggle to fully correct the errors in the base geometry. To address this issue, future research could explore end-to-end joint optimization strategies by integrating our module with foundation depth models. This would allow the explicit boundary priors extracted by $\alpha$Depth to back-propagate and iteratively correct global geometric distortions. Alternatively, introducing a confidence-aware fusion mechanism could enable the network to selectively rely on deep image semantics when the input depth exhibits low reliability.

**Two-Layer Representation**. Our approach models local soft boundaries based on the observation that a two-layer decomposition (foreground and background) is generally sufficient to resolve local occlusions. Thus, phenomena that are locally two-layer, including hair, fur, thin structures, and defocus or motion blur, can be well approximated by this formulation (see Fig. 14), as the alpha value corresponds to the point spread of defocus, the temporal coverage of motion, or the material transparency of the foreground layer. However, our approach may fail in cases that violate the local two-layer assumption. For example, volumetric media accumulate opacity continuously along the viewing ray and exhibit no well-defined boundary region. In addition, in some complex scenes where multiple semi-transparent boundaries overlap at the same pixel (*i.e.*, three or more overlapping layers, such as the intersecting points in the top example of Fig. 14), the current two-layer model might not fully capture all layered information. In future work, extending our framework to support an arbitrary number of overlapping layers presents an exciting research avenue. This could be achieved through an iterative layer-peeling mechanism, where foreground layers are sequentially stripped away [47]. Furthermore, integrating local volumetric representations (*e.g.*, localized radiance fields [28]) at intersecting boundaries could model complex semi-transparencies.

Table 15: **Ablation of multi-branch decoder** on matting datasets. The best results are marked.

| Method | AIM-500 | | | P3M-10K | | |
|---|---|---|---|---|---|---|
| | SAD ↓ | Grad ↓ | Conn ↓ | SAD ↓ | Grad ↓ | Conn ↓ |
| w/o Multi-Branch Decoder | 8.22 | 25.95 | 4.55 | 4.36 | 14.05 | 2.51 |
| **w/ Multi-Branch Decoder (Ours)** | 7.24 | 20.20 | 3.89 | 4.09 | 11.88 | 2.31 |

Table 16: **Ablation of edge extraction and semantic encoder** on Marvel-10K [62]. The best and second best results are marked.

| Method | Marvel-10K | | | |
|---|---|---|---|---|
| | S-PSNR ↑ | S-SSIM ↑ | LPIPS ↓ | DISTS ↓ |
| w/o Edge Extraction | 26.14 | 0.6718 | 0.1337 | 0.0910 |
| w/o Semantic Encoder | 27.35 | 0.7147 | 0.1269 | 0.0839 |
| **Full Model (Ours)** | 28.68 | 0.7636 | 0.1190 | 0.0738 |

**Video Consistency**. Although the proposed Circular Alpha Representation (CAR) demonstrates remarkable temporal consistency when processing video inputs (*e.g.*, see Fig. 9), the current $\alpha$Depth framework operates fundamentally as an image-based model. Due to the lack of explicit temporal constraints, $\alpha$Depth may produce results with flickering artifacts, especially in dynamic scenes where depth changes rapidly. Specifically, because $\alpha$Depth prioritizes capturing soft boundaries at depth discontinuities, it may fail to resolve boundaries at low depth gradients (*e.g.*, when two targets move close together in depth). Future extensions of this work could explore integrating spatio-temporal modules, *e.g.*, spatio-temporal attention, to further stabilize the layered predictions across video sequences.

# E Discussion of Societal Impacts

Our work on high-fidelity stereo conversion presents several positive societal impacts, primarily by democratizing the creation of immersive 3D content. By automating soft boundary decomposition in complex scenes, our $\alpha$Depth framework significantly lowers the barrier for creators in the Virtual and Augmented Reality (VR/AR), immersive education, and entertainment industries. This enables the efficient and low-cost transformation of legacy monocular media into engaging 3D experiences. However, we also acknowledge potential negative impacts associated with this technology. The ability to synthesize highly realistic stereo views could be misused to create immersive 3D disinformation, making fabricated content appear more physically credible. Additionally, the unauthorized stereo conversion of individuals or private scenes from casually captured 2D photos could raise privacy concerns.

# F Visualization of Ablation Models

Fig. 13 provides qualitative results for the ablation models detailed in Tab. 4. Due to depth ambiguity at soft boundaries, the baseline model (A#1) often generates warped images exhibiting broken boundaries and flying pixels. Although incorporating alpha estimation improves the structure of soft boundaries, the alpha valley issue inherent to the vanilla alpha representation tends to degrade warping performance (*e.g.*, see the bottom region of A#2 in Fig. 13). By contrast, the proposed circular alpha representation circumvents this issue, effectively preserving image structures and soft boundary details (A#3). Finally, when combined with the layered representation, our $\alpha$Depth faithfully recovers background information at soft boundaries, achieving the best warping performance (A#4).

Furthermore, the bottom row of Fig. 13 illustrates the qualitative impacts of the different alpha matting strategies evaluated in Tab. 4b. When employing the vanilla alpha representation (B#1), the network suffers from the alpha valley issue, which leads to unstable predictions and noticeable artifacts when extracting complex structures. Relying solely on the $\mathcal{L}_1$ loss without the matting loss $\mathcal{L}_\text{m}$ (B#2) fails to adequately capture fine-grained details, resulting in blurred and degraded soft boundaries. Similarly, utilizing single-stage training without the subsequent global refinement (B#3) yields noisy alpha predictions. By contrast, our full model (B#4) effectively overcomes these limitations and extracts intricate soft boundary details.

# G Visualization of $\alpha$Depth Results

Fig. 14 illustrates the results estimated by $\alpha$Depth in typical stereo conversion scenarios. Our method demonstrates robust performance in challenging conditions, such as multiple targets with defocus blur (top example)

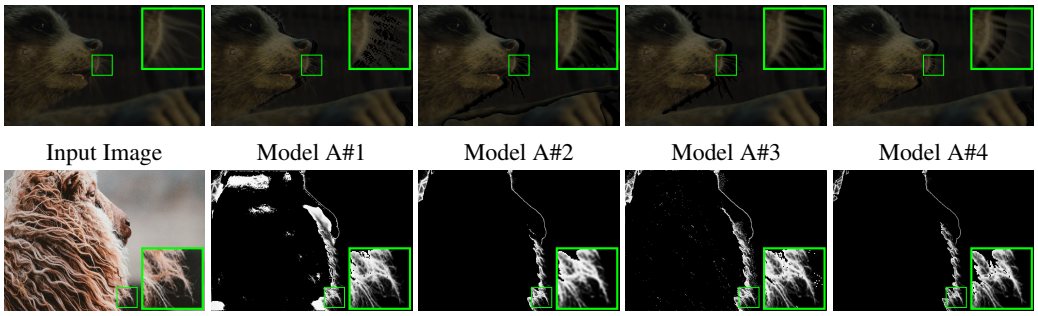



Figure 13: **Visual comparisons of ablation models in Tab. 4.**

and highly dynamic scenes with motion blur (middle example), highlighting its practical value for real-world applications.

# H More Visual Comparisons

## H.1 Stereo Conversion

In Figs. 15 and 16, we present additional qualitative comparisons evaluating warping and stereo conversion performance. When utilizing original depth maps estimated from Video Depth Anything [3], direct view transformation techniques struggle with the depth ambiguity at soft boundaries, frequently resulting in broken edges and flying pixels. While recent refinement approaches like HairGuard [62] capture finer boundary details, their reliance on a single-layer depth representation still leads to visible background bleeding and aliasing artifacts during warping. Furthermore, when comparing final stereo conversion results, existing state-of-the-art methods often struggle to maintain structural consistency at these intricate boundaries. In contrast, our $\alpha$Depth framework explicitly addresses these limitations by employing a layered representation that decouples soft boundaries into local foreground and background. By disentangling the mixed colors and resolving depth ambiguities, our method achieves superior fidelity in warping and stereo conversion results.

## H.2 Alpha Matting

Fig. 17 provides further visual comparisons of our Circular Alpha Representation (CAR) against state-of-the-art alpha matting baselines, including GVM [9], MatAnyone 2 [53], and ViTMatte [55]. Conventional matting techniques generally rely on explicit global definitions of foreground and background, necessitating manual guidance such as user-provided trimaps (*e.g.*, ViTMatte [55]) or segmentation masks (*e.g.*, MatAnyone 2 [53]) for instance-level inference. While auxiliary-free methods like GVM [9] reduce user effort, they are typically optimized for specific semantic categories and struggle to generalize to the diverse types of soft boundaries in complex scenes. As demonstrated in Fig. 17, our approach achieves comparable performance in extracting intricate soft boundary details with state-of-the-art matting methods, without any user intervention.

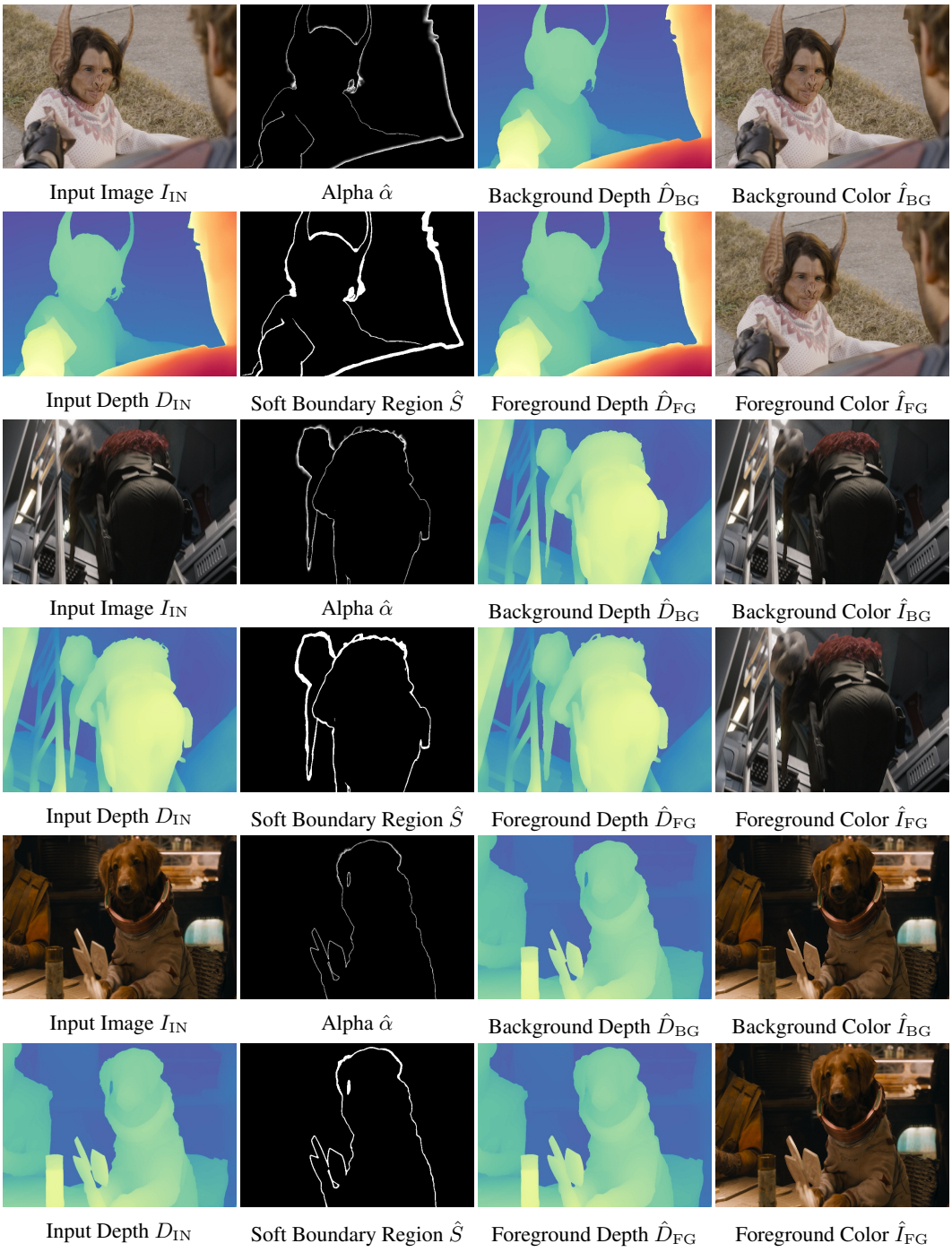



Figure 14: **Visualization of $\alpha$Depth results** on Marvel-10K dataset [62].

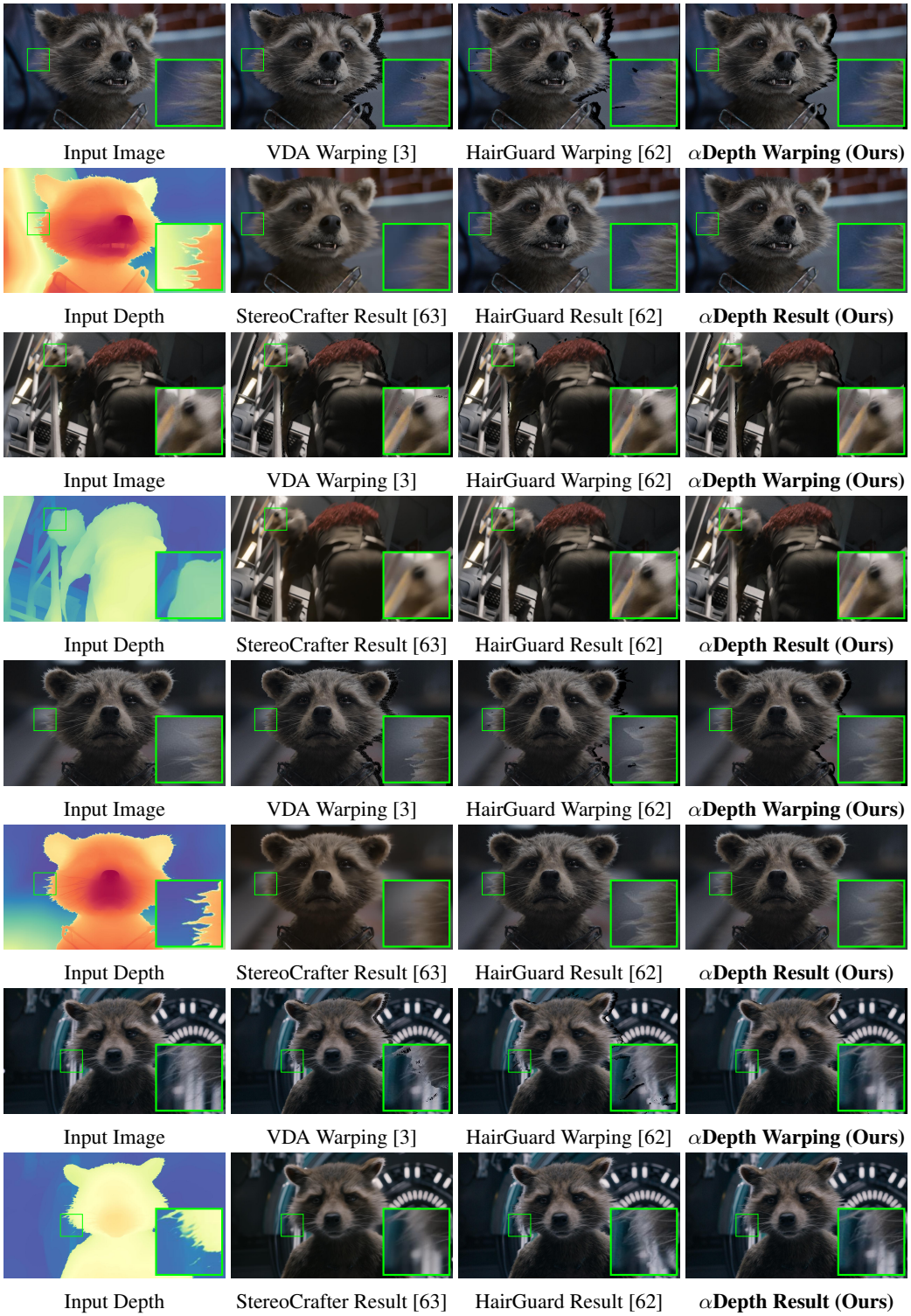



Figure 15: **Visual comparisons in warping and stereo conversion**, part one.




Figure 16: **Visual comparisons in warping and stereo conversion**, part two.




Figure 17: **Visual comparisons with alpha matting methods.**

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect our contribution.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see the Limitations and Discussions section (Sec. D) in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see the Experiments and Analysis section (Sec. 4).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This is research done in collaboration with a corporate research lab, and we haven't been able to get clearance to release the code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see the Experiments and Analysis section (Sec. 4) in the main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report a paired statistical analysis of our improvements over the strongest baseline (HairGuard) on both Mono2Stereo and Marvel-10K (Tab. 8). For each metric (PSNR, SSIM, LPIPS, DISTS), we provide (i) the mean per-sample difference with its 95% confidence interval obtained via percentile paired bootstrap (10,000 resamples with replacement over test samples), (ii) the one-sided Wilcoxon signed-rank p-value, which is distribution-free and does not assume normally distributed errors, and (iii) the per-sample win rate.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see the Experiments and Analysis section (experimental settings in Sec. 4).

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see the Discussion of Societal Impacts section (Sec. E) in the Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This work properly credits and respects the assets used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Please see the Method (Sec. 3) and the Experiments and Analysis section (Sec. 4).

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.