

Guardians of the Hair: Rescuing Soft Boundaries in Depth, Stereo, and Novel Views

Xiang Zhang^{1,2} Yang Zhang² Lukas Mehl² Markus Gross^{1,2} Christopher Schroers²
¹ETH Zürich ²DisneyResearch|Studios

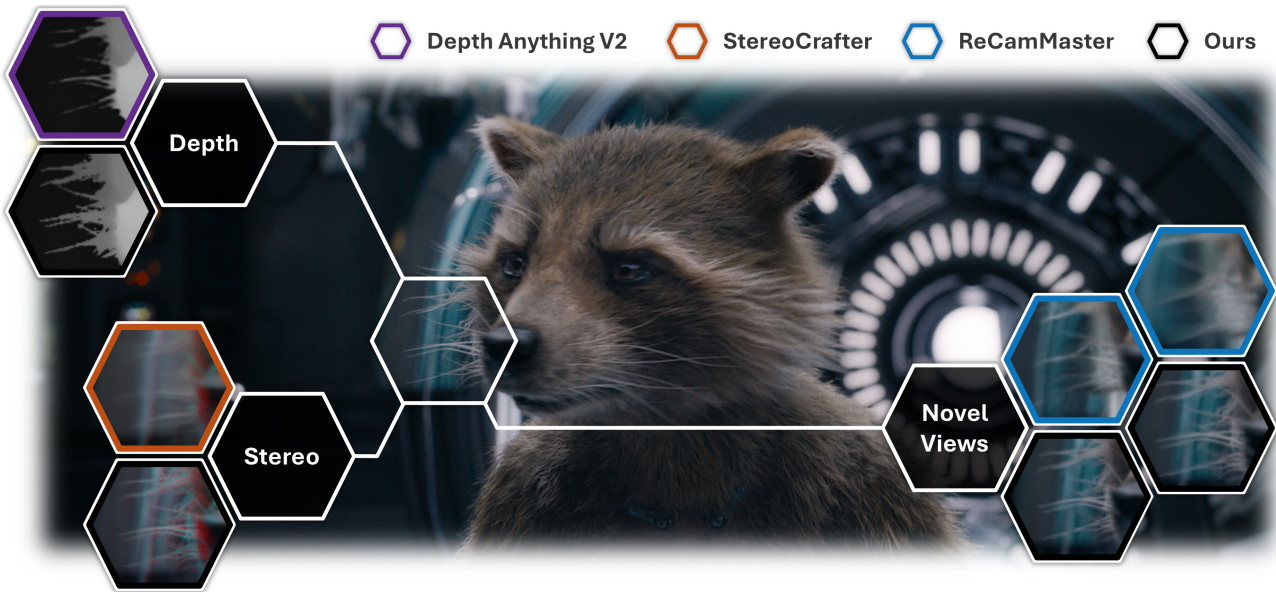


Figure 1. **Our Mission.** The Guardians of the Hair (**HairGuard**) aim to rescue soft boundary details, *e.g.*, thin hairs, where foreground and background are mixed in the observed color. Previous state-of-the-art approaches often suffer from missing details (see depth estimation results), degraded texture (see stereo results, displayed in anaglyph), and inconsistent geometry (see novel views) in soft boundaries. In contrast, HairGuard preserves fine-grained soft boundary details and demonstrates strong performance across diverse tasks.

Abstract

Soft boundaries, like thin hairs, are commonly observed in natural and computer-generated imagery, but they remain challenging for 3D vision due to the ambiguous mixing of foreground and background cues. This paper introduces **Guardians of the Hair (HairGuard)**, a framework designed to recover fine-grained soft boundary details in 3D vision tasks. Specifically, we first propose a novel data curation pipeline that leverages image matting datasets for training and design a depth fixer network to automatically identify soft boundary regions. With a gated residual module, the depth fixer refines depth precisely around soft boundaries while maintaining global depth quality, allowing plug-and-play integration with state-of-the-art depth models. For view synthesis, we perform depth-based forward warping

to retain high-fidelity textures, followed by a generative scene painter that fills disoccluded regions and eliminates redundant background artifacts within soft boundaries. Finally, a color fuser adaptively combines warped and inpainted results to produce novel views with consistent geometry and fine-grained details. Extensive experiments demonstrate that HairGuard achieves state-of-the-art performance across monocular depth estimation, stereo image/video conversion, and novel view synthesis, with significant improvements in soft boundary regions.

1. Introduction

Driven by recent advances in foundation models and large-scale visual datasets [44, 48], significant progress has been witnessed in the field of 3D vision, including depth es-

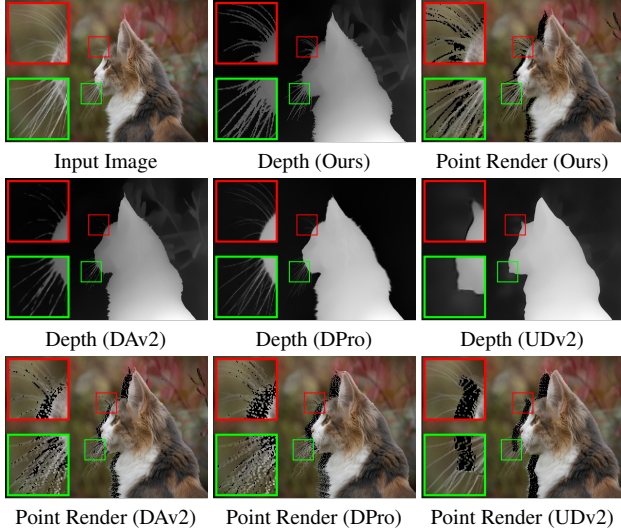


Figure 2. **Soft boundaries.** Existing depth estimation methods often struggle to capture accurate depth in soft boundaries, resulting in discontinuous depth (red box) and broken boundaries (green box). DAv2, DPro, and UDv2 represent Depth Anything V2 [67], Depth Pro [4], and UniDepthV2 [38], respectively.

timation [21, 42, 66, 77], stereo conversion [56, 73, 79], and novel view synthesis [2, 74, 78]. These techniques play a crucial role in understanding and reconstructing 3D scenes, with broad applications in robotics, autonomous driving, augmented/virtual reality (AR/VR), and film production [1, 11, 29, 33, 58]. Although existing methods have shown promising performance in general scenarios [2, 66], generating geometrically consistent and visually realistic results in scenes with soft boundaries, *e.g.*, hairs and thin structures, remains highly challenging (Fig. 1).

Soft boundaries are ubiquitous and often arise when pixels receive mixed contributions from both the foreground and background, due to thin/semi-transparent structures or alpha blending in rendering [28, 68]. Thus, they are commonly observed in natural images like shots containing animals and humans, as well as computer-generated imagery such as Fig. 1. The mixture of foreground and background pixels makes pixel-wise 3D estimation particularly challenging and ill-posed, since such regions exhibit uncertain correspondence and depth ambiguity.

Existing methods often struggle to capture accurate and fine-grained soft boundaries, as illustrated in Fig. 1. For example, the state-of-the-art monocular depth estimation method Depth Anything V2 [67] fails to extract the fine details of hairs and produces broken depth results (Fig. 1 and Fig. 2). Although the recent approach Depth Pro [4] achieves improved detail preservation in depth estimation, the estimated depth around soft boundaries often falls behind the true surface, leading to detached hairs as shown in the point cloud renders of Fig. 2 (red box). Since depth

estimation is often required by *explicit* 3D vision methods [73, 74, 78], the depth errors tend to propagate to the subsequent stages, resulting in sub-optimal performance. In the field of stereo conversion and novel view synthesis, one emerging trend is to generate new viewpoints in an *implicit* manner without depth [2, 11, 71]. By utilizing the rich prior knowledge learned in foundation generative models [44, 55], these implicit approaches can effectively handle complex occlusion and geometry in 3D world. However, due to the generative nature of the underlying foundation models, implicit 3D vision methods often suffer from hallucination issues and thus generate inconsistent texture details in soft boundaries (*e.g.*, see ReCamMaster [2] in Fig. 1). Meanwhile, most foundation generative models are designed in the latent space for computational efficiency [44, 55, 62]. Such a design often results in texture degradation due to pixel-to-latent compression [78], as illustrated by the StereoCrafter [79] results in Fig. 1.

In the realm of 2D vision, image matting provides an explicit formulation for soft boundaries by estimating an opacity map (*i.e.*, alpha matte) to model the pixel mixture along the transition between foreground and background [28, 68]. Inspired by the matting formulation, we leverage image matting datasets to improve soft boundary modeling and propose **Guardians of the Hair (HairGuard)** to rescue soft boundary details in 3D tasks. Specifically, HairGuard consists of three teammates: *depth fixer*, *scene painter*, and *color fuser*. By utilizing matting datasets in training data curation, our depth fixer learns to identify soft boundary regions and correct depth predictions with a gated residual module. This design not only enables precise depth refinement over soft boundaries, but also supports plug-and-play integration with zero-shot depth models for robust performance. For view synthesis, we first perform forward warping using the fixed depth, followed by a generative scene painter that synthesizes realistic disoccluded regions and corrects geometric errors caused by warping. Finally, to address texture hallucination and detail compression in generative models, we propose a color fuser to preserve fine-grained details and ensure geometrically consistent view synthesis via a dual skip module. As shown in Fig. 1, the components of HairGuard work collaboratively to achieve remarkable performance across different 3D vision tasks.

In a nutshell, our main contributions are three-fold:

- We present HairGuard to capture, model, and reconstruct fine-grained soft boundary details in 3D vision tasks. Extensive experiments verify the effectiveness and superiority of HairGuard across monocular depth estimation, stereo image/video conversion, and novel view synthesis.
- We design novel data curation strategies to leverage image matting datasets for training, enabling HairGuard to automatically identify and fix soft boundaries without relying on manually crafted cues, *e.g.*, trimaps [28, 68].

- We propose a depth fixer with a gated residual module, which enables precise depth refinement in soft boundary regions for plug-and-play enhancement. Additionally, we design a dual skip architecture in the color fuser to ensure geometrically consistent and high-quality view synthesis.

2. Related Work

2.1. Monocular Depth Estimation

Monocular depth estimation aims to infer scene geometry from a single image [6, 7, 18, 37, 43, 69, 70, 75], a fundamentally ill-posed problem due to the loss of depth cues during projection. To achieve zero-shot depth estimation, early attempts employ mixed training datasets to obtain a strong geometric prior of the scene [42, 43, 66]. Marigold also proposes to utilize the rich prior knowledge in generative foundation models, *e.g.*, Stable Diffusion [44], to efficiently approach zero-shot estimation [21]. Recently, several approaches have been proposed to improve the details of depth maps [4, 67, 77]. For example, Depth Anything V2 exploits high-quality depth supervision in synthetic datasets and learns to extract fine details from input images [67]. Meanwhile, Depth Pro designs a training protocol to combine real and synthetic datasets for metric depth estimation and fine boundary preservation. The recent approach UniDepthV2 also proposes an edge-guided loss to improve the sharpness of edges in the depth output [38]. Despite these advances, existing methods still struggle in soft boundary regions, often producing missing or discontinuous depth estimates. In contrast, our HairGuard precisely localizes soft boundaries and reconstructs fine-grained depth details, as shown in Fig. 2.

2.2. Stereo Conversion

The goal of stereo conversion is to generate right-view images/videos from left-view inputs [9, 33, 57, 61, 73], which has gained increasing attention due to its practical application in 3D video production and immersive media. With the rapid progress of generative foundation models [3, 44, 55], an emerging trend is to utilize learned generative and geometry priors for stereo conversion [13, 50, 73, 79]. For image-based conversion, StereoDiffusion introduces a training-free latent modification strategy using Stable Diffusion [56], and Mono2Stereo designs dual conditioning and edge-consistency losses to enhance stereo quality [73]. Recently, an increasing number of works have focused on leveraging video generative models for stereo video conversion [13, 50, 79]. For instance, StereoCrafter designs a tiled diffusion strategy to generate stereoscopic videos from high-resolution and long video inputs [79]. Based on Stable Video Diffusion [3], M2SVid devises a spatio-temporal aggregation mechanism to leverage information from neighboring frames and achieves high-quality inpainting perfor-

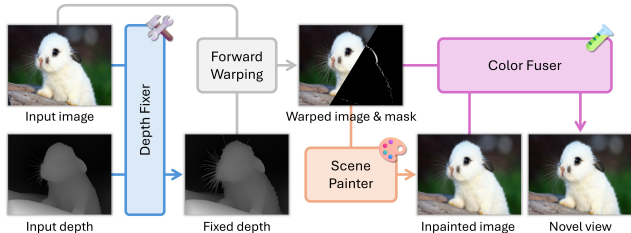


Figure 3. **HairGuard pipeline.** Given an input image and its estimated depth, we first design a depth fixer to refine depth predictions around soft boundary regions. The fixed depth is then used for forward warping to generate preliminary novel views, which are fed into the scene painter for disocclusion inpainting. Finally, our color fuser adaptively combines the warped and inpainted results to produce geometrically and visually consistent novel views.

mance [50]. Eye2Eye [13] further synthesizes stereo videos without explicit depth projection, effectively handling specular and transparent surfaces using video diffusion priors. However, due to the generative nature of diffusion models, current stereo conversion approaches often suffer from texture hallucination and loss of fine details, as shown in Fig. 1. To overcome these limitations, a color fuser network is designed in HairGuard to recover high-fidelity texture details.

2.3. Novel View Synthesis

Novel view synthesis has attracted considerable interest in computer vision community for its ability to render photo-realistic images from novel viewpoints [19, 22, 34, 52, 59, 60, 71, 74]. A popular trend is to perform 3D scene reconstruction from input images for novel view synthesis, such as Multi-Plane Image (MPI) [15, 25, 53], Neural Radiance Field (NeRF) [34, 72], and 3D Gaussian Splatting (3DGS) [22, 51, 64]. More recently, diffusion-based approaches have emerged as a powerful alternative, leveraging generative priors to produce high-fidelity novel views without requiring explicit 3D reconstruction [5, 14, 31, 46, 49, 80]. For example, ReCamMaster introduces frame-dimension conditioning to enhance view consistency in video diffusion models [2], but its results often suffer from texture inconsistency due to diffusion hallucination (*e.g.*, see Fig. 1). Another recent work, SplatDiff, integrates depth-guided pixel splatting with diffusion models to achieve high-fidelity view synthesis [78]. However, its performance is highly dependent on the quality of depth, which often contains errors around soft boundaries (Fig. 2). By comparison, our HairGuard combines a depth fixer and a color fuser to jointly correct depth inaccuracies and restore fine-grained texture details, achieving geometrically consistent and photo-realistic novel views (Fig. 1).

3. HairGuard

Following the formulation in image matting [28, 68], the observed image I can be expressed as an alpha composition

between the foreground I_{FG} and the background I_{BG} , *i.e.*,

$$I = \alpha \cdot I_{FG} + (1 - \alpha) \cdot I_{BG}, \quad (1)$$

where $\alpha \in [0, 1]$ denotes the opacity map (alpha matte). Soft boundaries can be defined as regions with mixed foreground and background pixels, *i.e.*, $\alpha \in (0, 1)$, posing ambiguity in depth and color correspondence. To handle these challenging areas in depth estimation (Sec. 3.1), we design a depth fixer to automatically localize soft boundaries and refine depth predictions, as shown in Fig. 3. For view synthesis tasks (Sec. 3.2), we first perform forward warping based on the fixed depth, and then apply the generative scene painter to fill the unknown regions like disoccluded areas. Finally, our color fuser combines warped and inpainted results for high-quality view synthesis.

3.1. Depth Estimation

Given an image and its depth map (*e.g.*, estimation results from Depth Anything V2 [67]), our depth fixer aims to automatically identify soft boundary regions and perform precise depth correction. However, several challenges exist:

- *High-quality annotation.* Most existing depth datasets focus on scenes with hard boundaries, lacking fine-grained depth annotations around soft boundary regions.
- *Automatic localization.* Estimation in soft boundaries often relies on hand-crafted cues like trimaps [68], hindering generalization and applicability to complex scenes.
- *Precise refinement.* Achieving precise depth correction in soft boundary regions without compromising the global depth quality remains an open challenge.

Dataset Curation. Collecting large-scale datasets with high-quality depth annotations in soft boundary regions could be time-consuming and impractical. Thus, we address the *high-quality annotation* issue by utilizing the existing image matting datasets, which contain diverse targets with soft boundaries and the corresponding opacity maps (*i.e.*, alpha mattes). As shown in Fig. 4a, we use matting datasets as foreground datasets $\mathcal{I}_{FG} = \{(\alpha, I_{FG})_i\}_i$ and image datasets as background datasets $\mathcal{I}_{BG} = \{I_{BG,i}\}_i$. Since alpha mattes usually exhibit smooth transitions in soft boundaries, which are not aligned with depth characteristics, we first obtain alpha masks M_α by thresholding α with α_{th} , *i.e.*, $M_\alpha = \{p \mid \alpha_{th} < \alpha(p)\}$. Then, we generate foreground depth d_{FG} by

$$d_{FG} = M_\alpha \odot \text{Depth}(I_{FG}), \quad (2)$$

where $\text{Depth}(\cdot)$ represents depth estimation methods [67], and a green background is added to I_{FG} to enhance the contrast in depth estimation. Afterward, we obtain the background depth as $d_{BG} = \text{Depth}(I_{BG})$ and randomly sample two depth values from $[d_{min}, d_{max}]$ to rescale d_{FG} for data augmentation, where $d_{min} = \max_{p \in M_\alpha} (d_{BG}(p))$ to ensure

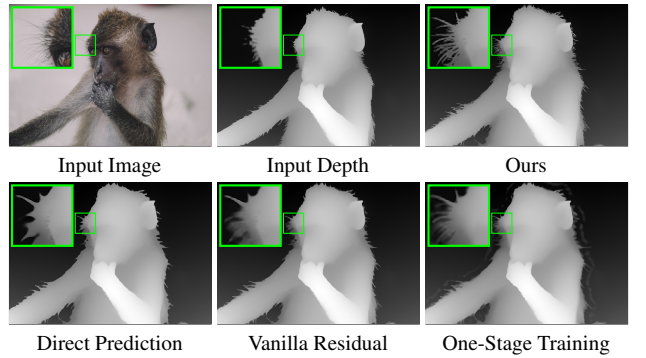
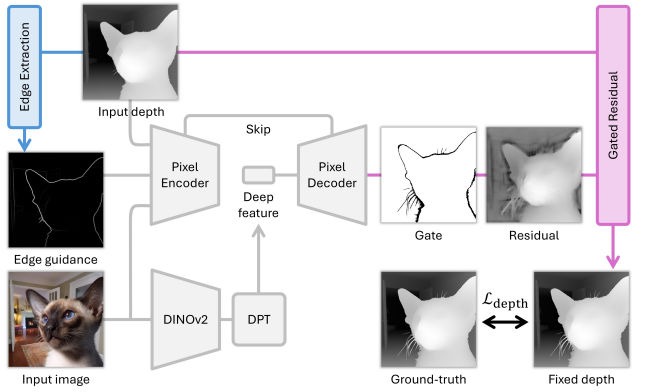
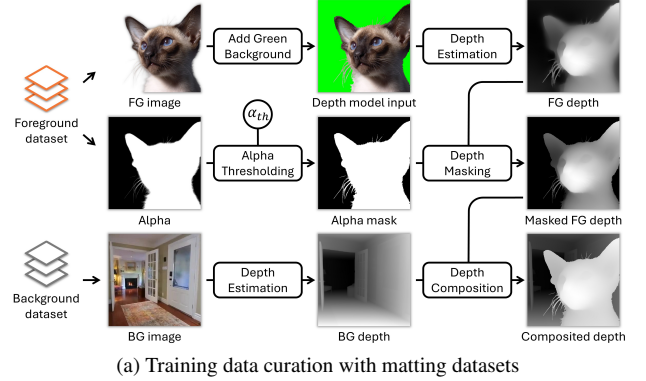


Figure 4. **Depth fixer.** (a) We utilize image matting datasets to synthesize training data with fine-grained depth labels in soft boundaries. (b) Instead of relying on manually crafted cues like trimaps [68], we leverage depth maps and image semantics to automatically identify soft boundary regions. The gated residual module enables precise depth correction in soft boundary areas and thus benefits plug-and-play refinement. (c) Compared with direct prediction and vanilla residual, our gated residual combined with two-stage training achieves the best depth results.

correct depth ordering and d_{max} is a predefined constant. Finally, we blend d_{FG} and d_{BG} by depth composition:

$$d = d_{FG} \odot M_\alpha + d_{BG} \odot (1 - M_\alpha). \quad (3)$$

Using Eq. (3), one can create depth training pairs $\{(d_{in}, d_{GT})_i\}_i$ for the depth fixer by varying the threshold

α_{th} . A lower α_{th} is used to generate depth labels d_{GT} with fine details in soft boundaries, and a higher α_{th} is used to simulate depth inputs d_{in} with broken or missing depth in these regions. Additionally, we apply a random Gaussian blur to M_α when generating d_{in} , but use the unblurred mask in Eq. (3) to produce d_{GT} with sharp boundaries.

Network Design. As illustrated in Fig. 4b, our depth fixer has two main branches: a feature branch built upon DI-NOv2 [36] and DPT [43] to extract deep features and image semantics, and a pixel branch based on U-Net [45] to capture local structures and boundary details. To address the *automatic localization* problem, we propose to infer soft boundaries directly from images and depth maps. In particular, we first generate explicit edge guidance e by applying the Sobel operator to the input depth, *i.e.*, $e = \text{Sobel}(d_{in})$, and then concatenate e with image I_{in} and depth d_{in} as inputs to the pixel branch. With this design, our depth fixer is able to focus on the regions with high depth gradients and learn to automatically identify soft boundaries with image semantics and geometric layouts.

For *precise refinement*, we propose a gated residual mechanism to refine depth only in soft boundary regions while preserving global depth quality. Specifically, we first model the soft boundary regions by predicting a gate map $G \in [0, 1]$, where $G < 1$ indicates soft boundary regions. Then, the gated residual is performed to obtain the refined depth \hat{d} ,

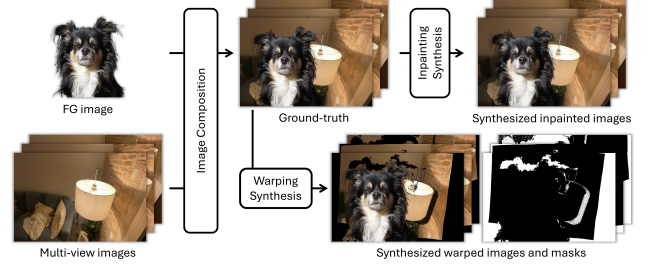
$$\hat{d} = d_{in} \cdot G + d_{res} \cdot (1 - G), \quad (4)$$

where d_{res} is the estimated depth residual. Compared with the direct prediction of refined depth or vanilla residual approach, our gated residual better preserves sharp and fine-grained details in soft boundaries, as shown in Fig. 4c. Furthermore, the gating mechanism decouples depth estimation and soft-boundary fixing, and thus our depth fixer can be seamlessly integrated with state-of-the-art depth models to achieve robust and detail-preserving performance.

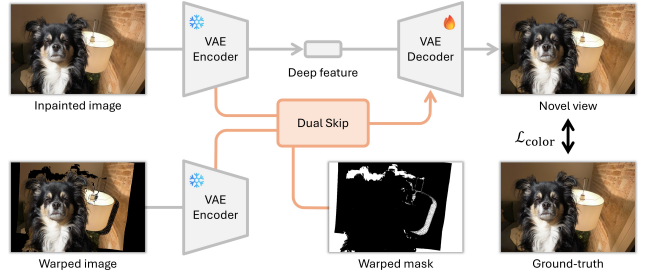
Model Training. Directly training the depth fixer using standard depth losses [42] tends to yield a trivial solution where the gate collapses to $G = 1$. Hence, we propose a two-stage strategy to learn depth refinement in a local-to-global manner. We first generate a soft boundary mask M_{soft} by thresholding the ground-truth alpha matte, *i.e.*, $M_{soft} = \{p \mid \alpha_{min} < \alpha(p) < \alpha_{max}\}$, with constants $\alpha_{min}, \alpha_{max}$ determining the soft boundary areas. The learning objective for the first stage $\mathcal{L}_{depth}^{stage1}$ is defined as

$$\mathcal{L}_{depth}^{stage1} = \mathcal{L}_1(\hat{d}, d_{GT}) + \mathcal{L}_\alpha(\hat{d} \odot M_{soft}, d_{GT} \odot M_{soft}), \quad (5)$$

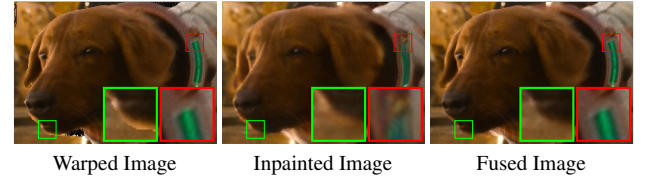
where \mathcal{L}_1 denotes the ℓ_1 loss, and \mathcal{L}_α is an image matting loss from ViTMatte [68] to facilitate detail extraction. Although $\mathcal{L}_{depth}^{stage1}$ prevents the trivial solution of $G = 1$ by imposing stronger penalties on soft boundaries, it often introduces halo artifacts around these regions, as illustrated in



(a) Training data curation with matting datasets



(b) Network architecture



(c) Comparisons of warped, inpainted, and fused images

Figure 5. **Color fuser.** (a) We employ image matting datasets and multi-view datasets to synthesize warped and inpainted results for training. (b) Built upon a pre-trained VAE, we design a dual skip module to leverage the merits of inpainted and warped images. (c) Our color fuser eliminates redundant background colors (green box in the warped image) and hallucinated textures (red box in the inpainted image) for high-quality view synthesis.

Fig. 4c. Thus, in the second stage, we apply the constraint \mathcal{L}_α globally to improve the overall depth quality, *i.e.*,

$$\mathcal{L}_{depth}^{stage2} = \mathcal{L}_\alpha(\hat{d}, d_{GT}). \quad (6)$$

Fig. 4c shows that our two-stage training achieves fine-grained details while preserving global depth quality.

3.2. View Synthesis

For view synthesis, we first perform forward warping based on the fixed depth to preserve fine details and soft boundaries. However, the mixture of foreground and background in soft boundaries often introduces redundant background colors in the warped results (*e.g.*, see the green box in Fig. 5c). Since existing multi-view datasets mainly contain hard boundaries, we propose to model such characteristics by leveraging matting datasets in data curation (Fig. 5a).

Dataset Curation. Given a sequence of background images $\{I_{BG,i}\}_i$ from multi-view datasets, we first predict the optical flow $\{f_{BG,i}\}_i$ between all pairs of images via an

Table 1. **Zero-shot depth boundary accuracy** on the natural image matting datasets. Depth fixer can be integrated with different depth models in a plug-and-play manner for soft boundary refinement. **Best** results are marked. Please see the supplementary for more evaluation.

Method	AIM-500				P3M-10K			
	DBE_comp ↓	DBE_acc ↓	EP (%) ↑	ER (%) ↑	DBE_comp ↓	DBE_acc ↓	EP (%) ↑	ER (%) ↑
Depth Anything V2 [67]	7.93	3.29	19.90	6.50	7.53	2.60	26.53	9.37
Depth Anything V2+Depth Fixer (Ours)	7.19	2.10	34.56	13.08	7.21	1.93	36.91	13.39
Depth Pro [4]	7.75	3.80	15.92	6.12	7.25	3.25	18.36	9.21
Depth Pro+Depth Fixer (Ours)	6.70	2.30	35.01	17.33	6.44	1.78	37.90	18.91
UniDepthV2 [38]	8.34	3.87	19.52	5.14	7.73	3.48	20.82	8.12
UniDepthV2+Depth Fixer (Ours)	7.49	2.71	33.06	10.98	6.89	2.05	37.71	15.12

off-the-shelf optical flow estimator [63]. To synthesize the warped results of soft boundaries, we sample a foreground image I_{FG} from the matting dataset, and generate the foreground flow f_{FG} with a random displacement vector (u, v) for all pixels, ensuring purely translational motion within the image plane. Then, we perform flow composition using alpha mask M_α ,

$$f = f_{FG} \odot M_\alpha + f_{BG} \odot (1 - M_\alpha). \quad (7)$$

Since the foreground only moves within the image plane, ground-truth views $\{I_{GT,i}\}_i$ can be easily synthesized by applying Eq. (1) to the foreground I_{FG} and background images $\{I_{BG,i}\}_i$. Although the foreground motions are relatively simple, the background regions preserve realistic viewpoint changes and complex camera motions for robust training. Finally, we perform forward warping with flows $\{f_i\}_i$ to generate the warped images and masks, and fine-tune our scene painter to fit the characteristics of soft boundaries while inpainting disoccluded regions. The aligned synthesis strategy in SplatDiff [78] is also applied to the background regions for precise viewpoint control.

Color Fuser. Although the scene painter is able to eliminate redundant background in soft boundaries, its generative nature tends to hallucinate inconsistent texture details (*e.g.*, see the red box in Fig. 5c). To this end, we propose the color fuser to adaptively combine warped and inpainted images. As shown in Fig. 5b, we build the color fuser upon a pre-trained Variational Auto-Encoder (VAE) [39] to harness its reconstruction prior. Since VAE models often suffer from detail compression [78], a dual skip module is designed to propagate fine-grained features for fusion. Specifically, we first extract multi-scale features of the inpainted and warped images via a frozen VAE encoder. These features are then concatenated with the warped masks and fed into the VAE decoder to compensate for texture details. Based on our curated view synthesis dataset, we further synthesize input inpainted images with hallucinated textures by applying the scene painter to ground-truth images $\{I_{GT,i}\}_i$. Finally, we fine-tune the VAE decoder using the following objective:

$$\mathcal{L}_{color} = \mathcal{L}_1(\hat{I}, I_{GT}) + \lambda \cdot \mathcal{L}_{lpips}(\hat{I}, I_{GT}), \quad (8)$$

where the balancing parameter $\lambda = 0.1$, \hat{I} denotes the

outputs of the color fuser, and \mathcal{L}_{lpips} indicates perceptual loss [76]. Compared with the warped and inpainted results, our color fuser produces the best novel views with high-quality texture and geometry (Fig. 5c).

4. Experiments and Analysis

4.1. Experimental Settings

Implementation Details. For depth fixer, we curate our training dataset with $\alpha_{min} = 0.02$ and $\alpha_{max} = 0.98$. We use $\alpha_{th} = \alpha_{min}$ when generating ground-truth depth, and randomly sample $\alpha_{th} \sim \mathcal{U}(\alpha_{min}, \alpha_{max})$ when synthesizing input depth. We implement the depth fixer with Depth Anything V2 [67] weight initialization for the feature branch. Depth fixer is trained with AdamW optimizer [32] under 448×448 patches, batch size 32, and 1×10^{-5} learning rate for 35K iterations for both stages. For scene painter, we employ the pretrained VACE model [20] based on Wan2.1-1.3B [55], and fine-tune it under 480×832 resolution, batch size 4, and 1×10^{-5} learning rate for 10K iterations. Regarding the color fuser, we add extra residual blocks in VAE decoder to blend the features from dual skip module, and fine-tune under 448×448 patches, batch size 16, and 1×10^{-5} learning rate for 35K iterations. The total training takes 4 days on 4 NVIDIA RTX A6000 GPUs.

Datasets. For training, we employ two multi-view datasets as background datasets: *RealEstate10K* [81] and *DL3DV-10K* [30], and three image matting datasets as foreground datasets: *AM-2K* [24], *Distinctions-646* [40], and *Composition-1K* [65]. For evaluation, we created a *Marvel-10K* dataset composed of 501 stereo videos from Marvel movies, with a total of 12,525 stereo pairs. We also use 5 public depth estimation benchmarks for zero-shot evaluation: *NYUV2* [35], *KITTI* [12], *ETH3D* [47], *ScanNet* [8], and *DIODE* [54]. In addition, two natural image matting datasets *AIM-500* [27] and *P3M-10K* [26] are employed to evaluate the real-world performance of HairGuard.

4.2. Depth Estimation

We apply depth fixer to improve 3 state-of-the-art depth estimation models: Depth Anything V2 [67], Depth Pro [4], and UniDepthV2 [38] in a plug-and-play manner, and eval-

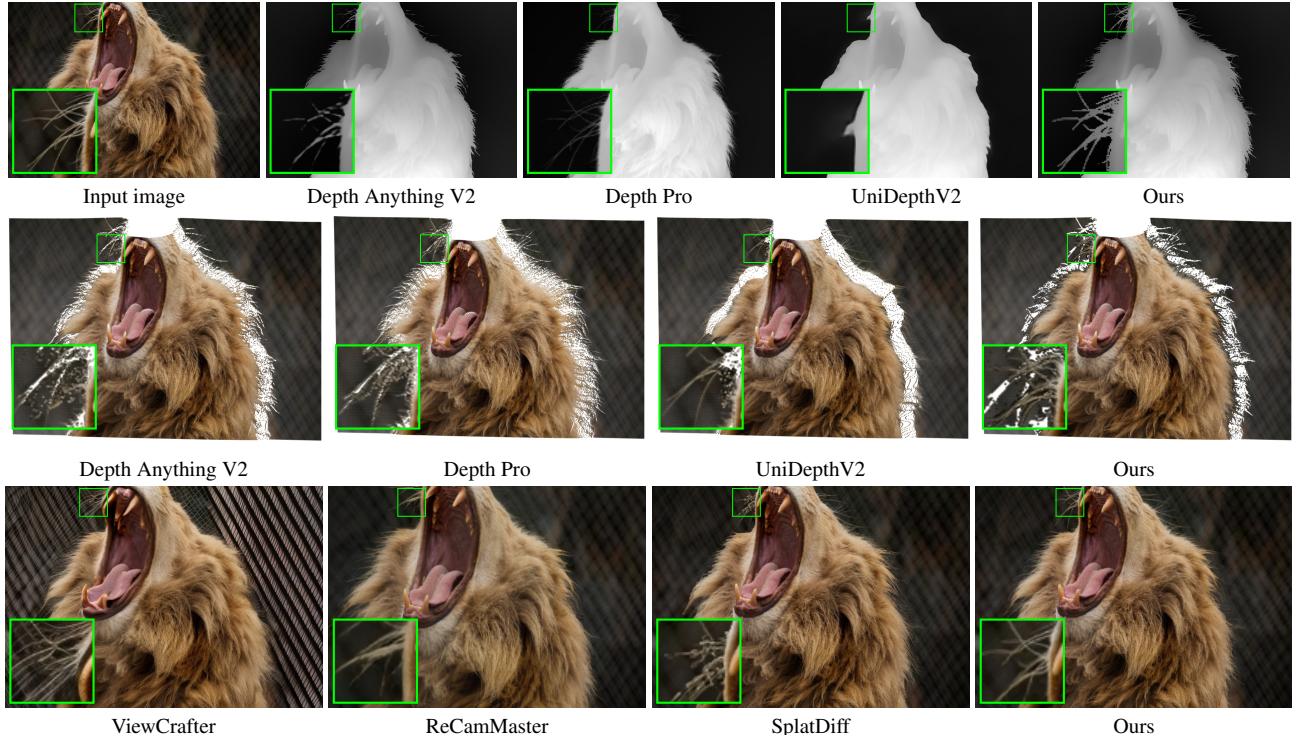


Figure 6. **Qualitative comparisons** of depth estimation (top), point clouds (middle), and novel view synthesis (bottom). Our HairGuard better preserves soft boundary details in depth results and novel views, without artifacts like broken, detached, or hallucinated hairs.

Table 2. **Zero-shot depth estimation performance.** The depth fixer preserves the zero-shot capability of its base depth model in diverse scenarios. Metrics are shown in percentage. **Better** results are marked. Please see the supplementary for more robustness evaluations.

Method	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑	AbsRel ↓	$\delta 1$ ↑
Depth Anything V2 [67]	4.27	97.86	7.97	94.38	5.25	98.27	4.15	97.94	26.24	75.49
Depth Anything V2+Depth Fixer (Ours)	4.27	97.86	7.97	94.38	5.25	98.27	4.15	97.94	26.24	75.49
Depth Pro [4]	4.29	97.90	5.98	96.25	5.23	96.89	4.11	97.98	22.20	76.28
Depth Pro+Depth Fixer (Ours)	4.29	97.90	5.98	96.25	5.23	96.89	4.11	97.98	22.17	76.28
UniDepthV2 [38]	3.40	98.33	4.67	97.42	3.31	99.16	3.08	98.32	23.94	75.67
UniDepthV2+Depth Fixer (Ours)	3.40	98.33	4.67	97.42	3.27	99.16	3.08	98.32	23.87	75.71

uate their performance in terms of depth boundary accuracy and zero-shot depth estimation.

Boundary Accuracy. Following Depth Pro [4], we employ image matting datasets to evaluate depth accuracy in soft boundaries. Edge-based metrics, *i.e.*, the completeness and accuracy of depth boundaries (DBE_comp and DBE_acc) [23] and the edge precision and recall (EP and ER) [17], are used to evaluate depth results in soft boundary regions. Our depth fixer better captures fine-grained depth details in soft boundaries (see depth and point cloud results in Fig. 6), and consistently yields significant improvements when integrated with different depth models (Tab. 1).

Zero-Shot Performance. We further test the robustness of the depth fixer on 5 unseen public datasets. Although these datasets rarely contain soft boundaries, Tab. 2 shows that our depth fixer still achieves comparable or slightly better performance under in-the-wild settings. Thanks to the pro-

posed gated residual module, our depth fixer can adaptively fix depth in soft boundaries while maintaining the zero-shot performance of the base depth model, allowing seamless plug-and-play integration with current and future state-of-the-art depth models.

4.3. Stereo Conversion

Since stereo conversion is widely applied in film production, we compare HairGuard with state-of-the-art stereo conversion and novel view synthesis approaches on the Marvel-10K dataset, which features challenging cinematic scenes and talking heads with complex hair structures. Pixel-level metrics (PSNR, SSIM, and RMSE), feature-level metrics (LPIPS [76] and DISTS [10]), and the stereo metric SIoU [73] are used for evaluation.

Benchmarking on Marvel-10K. We compare the performance of stereo image conversion (1 frame per sequence)

Table 3. **Stereo image/video conversion performance** on the Marvel-10K dataset. The **best** and **second-best** results are marked.

Method	Stereo Image Conversion						Stereo Video Conversion					
	PSNR \uparrow	SSIM \uparrow	RMSE \downarrow	LPIPS \downarrow	DISTS \downarrow	SIoU \uparrow	PSNR \uparrow	SSIM \uparrow	RMSE \downarrow	LPIPS \downarrow	DISTS \downarrow	SIoU \uparrow
StereoDiffusion [56]	32.70	0.7654	6.05	0.2177	0.0698	0.2638	32.71	0.7656	6.04	0.2172	0.0693	0.2655
Mono2Stereo [73]	33.65	0.8143	5.45	0.1973	0.0690	0.2556	33.63	0.8134	5.47	0.1980	0.0691	0.2552
StereoCrafter [79]	32.52	0.8148	6.13	0.2330	0.1208	0.2664	32.35	0.8125	6.25	0.2381	0.1246	0.2645
ViewCrafter [74]	30.69	0.6705	7.61	0.3258	0.1330	0.2085	30.73	0.6739	7.59	0.3221	0.1312	0.2101
NVS-Solver [71]	31.18	0.7108	7.16	0.3323	0.1793	0.2143	31.44	0.7220	6.98	0.3256	0.1743	0.2161
ReCamMaster [2]	30.44	0.6118	7.82	0.4082	0.1391	0.1798	30.41	0.6107	7.84	0.4106	0.1412	0.1797
SplatDiff [78]	36.23	0.8857	4.06	0.1116	0.0435	0.3259	36.24	0.8858	4.06	0.1114	0.0437	0.3280
HairGuard (Ours)	36.59	0.8953	3.91	0.0909	0.0331	0.3337	36.58	0.8953	3.92	0.0911	0.0334	0.3355

and stereo video conversion (all sequence frames) in Tab. 3. Although our HairGuard focuses mainly on improving soft boundaries, which usually occupy small regions in images, it consistently outperforms existing approaches in all metrics. In addition, Fig. 7 verifies the superior performance and temporal consistency of HairGuard compared with previous video-based methods.

Ablation Study. Tab. 4 shows the contribution of each component in our HairGuard. We first estimate depth with Depth Anything V2 [67] and use its warped results as the baseline (#1). By fixing depth in soft boundaries, depth fixer achieves better warping performance with higher SiOU (#2 vs. #1). Scene painter largely improves perceptual quality by filling disoccluded regions, but suffers from detail compression and texture hallucination (better LPIPS and worse PSNR in #3). By adaptively combining warped and in-painted images via the color fuser, HairGuard achieves the best results with high-quality textures and stereo effects.

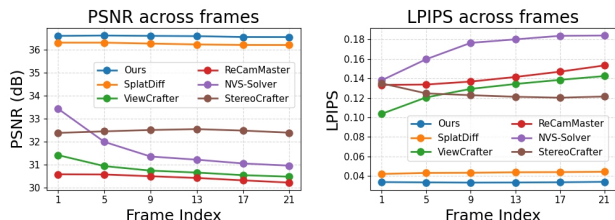


Figure 7. **Stereo video conversion performance** on Marvel-10K.

Table 4. **Ablation on stereo image conversion** on the Marvel-10K dataset. The **best** and **second-best** results are marked. Please see the supplementary for more detailed ablation studies.

Exp	Depth Fixer	Scene Painter	Color Fuser	Marvel-10K		
				PSNR \uparrow	LPIPS \downarrow	SIoU \uparrow
#1				36.26	0.1490	0.3097
#2	✓			36.28	0.1458	0.3118
#3	✓	✓		35.82	0.1246	0.3015
#4	✓	✓	✓	36.59	0.0909	0.3337

4.4. Novel View Synthesis

Benchmarking on Matting Datasets. We employ 2 natural image matting datasets to evaluate the novel view synthesis performance on scenes with soft boundaries. Since ground-truth views are not available, we adopt FID [16] and the av-

Table 5. **Novel view synthesis performance** on the natural image matting datasets. The **best** and **second-best** results are marked.

Method	AIM-500		P3M-10K	
	FID \downarrow	CLIP-F \uparrow	FID \downarrow	CLIP-F \uparrow
NVS-Solver [71]	51.71	97.24	55.12	96.66
ViewCrafter [74]	33.43	99.03	35.40	98.73
ReCamMaster [2]	57.19	97.95	62.80	97.09
SplatDiff [78]	19.26	99.36	21.61	99.09
HairGuard (Ours)	18.82	99.38	21.38	99.11

erage CLIP similarity of adjacent frames (CLIP-F) [41] for quantitative evaluation. Fig. 6 and Tab. 5 verify the state-of-the-art performance of HairGuard in real-world scenarios.

User Study. We conducted a user study with 27 participants on the full evaluation sets of AIM-500 and P3M-10K datasets (1000 natural images in total, no hand-picked samples). The participants will see side-by-side novel view video results, vote for their preferred one, and indicate if it is a strong preference. 1332 votes are collected in total, and the results in Fig. 8 verify the superiority of HairGuard.

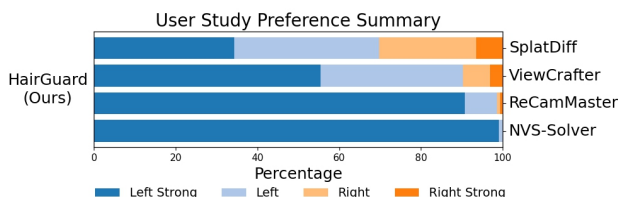


Figure 8. **User study on novel view synthesis.** A survey of 27 participants (1332 votes in total) shows the superiority of our HairGuard compared with previous state-of-the-art approaches.

5. Conclusion

This paper presents HairGuard to address the challenges of soft boundaries in 3D vision tasks. By utilizing image matting datasets, we train a depth fixer to automatically identify soft boundary regions and correct depth results in a plug-and-play manner. For view synthesis tasks, the scene painter and color fuser are employed to fix geometric errors in novel views while preserving high-quality texture details. Extensive experiments on monocular depth estimation, stereo image/video conversion, and novel view synthesis verify the state-of-the-art performance of HairGuard.

References

- [1] Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025. 2
- [2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuoqiu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *ICCV*, 2025. 2, 3, 8
- [3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3
- [4] 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. 2, 3, 6, 7
- [5] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *ICCV*, pages 4217–4229, 2023. 3
- [6] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NIPS*, 2016. 3
- [7] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *CVPR*, pages 679–688, 2020. 3
- [8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 6
- [9] Peng Dai, Feitong Tan, Qiangeng Xu, David Futschik, Ruofei Du, Sean Fanello, Xiaojuan Qi, and Yinda Zhang. Svc: 3d stereoscopic video generation via denoising frame matrix. *arXiv preprint arXiv:2407.00367*, 2024. 3
- [10] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2022. 7
- [11] 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. 2
- [12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 6
- [13] 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. 3
- [14] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. *arXiv preprint arXiv:2501.03847*, 2025. 3
- [15] Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*, New York, NY, USA, 2022. Association for Computing Machinery. 3
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 8
- [17] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *WACV*, pages 1043–1051. IEEE, 2019. 7
- [18] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *PAMI*, 2024. 3
- [19] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. Slide: Single image 3d photography with soft layering and depth-aware inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12518–12527, 2021. 3
- [20] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 6
- [21] 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. 2, 3
- [22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 3
- [23] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *ECCVW*, 2018. 7
- [24] Jizhizi Li. *End-to-end Animal Matting*. PhD thesis, University of Sydney, 2020. 6
- [25] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *ICCV*, pages 12578–12588, 2021. 3
- [26] Jizhizi Li, Sihan Ma, Jing Zhang, and Dacheng Tao. Privacy-preserving portrait matting. In *ACMMM*, pages 3501–3509, 2021. 6
- [27] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep automatic natural image matting. In *IJCAI*. International Joint Conferences on Artificial Intelligence Organization, 2021. 6
- [28] Jizhizi Li, Jing Zhang, and Dacheng Tao. Deep image matting: A comprehensive survey. *arXiv preprint arXiv:2304.04672*, 2023. 2, 3
- [29] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Pla-

- taniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. *arXiv preprint arXiv:2412.12091*, 2024. 2
- [30] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, pages 22160–22169, 2024. 6
- [31] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, pages 9298–9309, 2023. 3
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6
- [33] Lukas Mehl, Andrés Bruhn, Markus Gross, and Christopher Schroers. Stereo conversion with disparity-aware warping, compositing and inpainting. In *WACV*, pages 4260–4269, 2024. 2, 3
- [34] 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. 3
- [35] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 6
- [36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafranec, 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. 5
- [37] 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. 3
- [38] 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. 2, 3, 6, 7
- [39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 6
- [40] 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. 6
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmlR, 2021. 8
- [42] 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, 2020. 2, 3, 5
- [43] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 3, 5
- [44] 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. 1, 2, 3
- [45] 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. 5
- [46] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *CVPR*, pages 9420–9429, 2024. 3
- [47] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017. 6
- [48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NIPS*, pages 25278–25294, 2022. 1
- [49] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. In *NeurIPS*, 2024. 3
- [50] 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. 3
- [51] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024. 3
- [52] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *CVPR*, pages 10208–10217, 2024. 3
- [53] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, pages 551–560, 2020. 3
- [54] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *arXiv preprint arXiv:1908.00463*, 2019. 6
- [55] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video gen-

- erative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 6
- [56] 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. 2, 3, 8
- [57] Xiaodong Wang, Chenfei Wu, Shengming Yin, Minheng Ni, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Fan Yang, Lijuan Wang, Zicheng Liu, et al. Learning 3d photography videos via self-supervised diffusion on single images. *arXiv preprint arXiv:2302.10781*, 2023. 3
- [58] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, pages 8445–8453, 2019. 2
- [59] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *CVPR*, pages 7467–7477, 2020. 3
- [60] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *CVPR*, pages 9076–9086, 2023. 3
- [61] Junyuan Xie, Ross Girshick, and Ali Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *European conference on computer vision*, pages 842–857. Springer, 2016. 3
- [62] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *ECCV*, pages 399–417. Springer, 2025. 2
- [63] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofghi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE TPAMI*, 2023. 6
- [64] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthspat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 3
- [65] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2970–2979, 2017. 6
- [66] 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. 2, 3
- [67] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 2, 3, 4, 6, 7, 8
- [68] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pre-trained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 2, 3, 4, 5
- [69] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, pages 204–213, 2021. 3
- [70] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, pages 9043–9053, 2023. 3
- [71] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In *ICLR*, 2025. 2, 3, 8
- [72] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 3
- [73] 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. 2, 3, 7, 8
- [74] 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. 2, 3, 8
- [75] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. In *NIPS*, pages 14128–14139, 2022. 3
- [76] 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. 6, 7
- [77] 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. 2, 3
- [78] Xiang Zhang, Yang Zhang, Lukas Mehl, Markus Gross, and Christopher Schroers. High-fidelity novel view synthesis via splatting-guided diffusion. In *SIGGRAPH*, New York, NY, USA, 2025. Association for Computing Machinery. 2, 3, 6, 8
- [79] 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. 2, 3, 8
- [80] Chuanxia Zheng and Andrea Vedaldi. Free3d: Consistent novel view synthesis without 3d representation. In *CVPR*, pages 9720–9731, 2024. 3
- [81] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), 2018. 6