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

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

Abstract

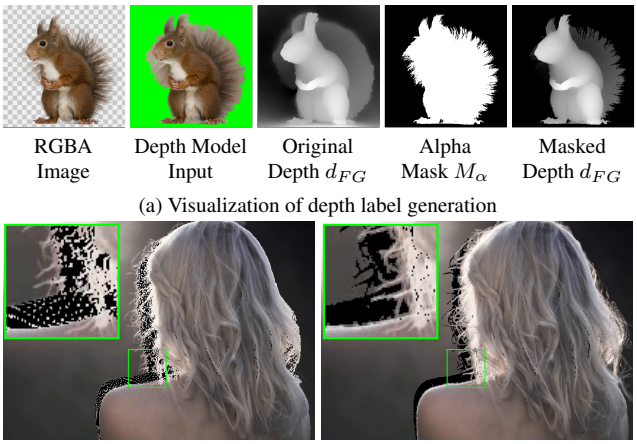
The supplementary material is organized as follows: We first provide more implementation details in Sec. 1. Then, the model performance, including robustness, plug-and-play performance, and computational complexity, is analyzed in Sec. 2. Following that, we show more experiments on the depth fixer and the color fuser in Sec. 3 and Sec. 4, respectively. Afterward, we analyze the limitations of HairGuard and discuss potential future directions in Sec. 5. In the end, more visual comparisons on ablation study, monocular depth estimation, stereo conversion, and novel view synthesis are provided in Sec. 6.

1. More Implementation Details

1.1. Marvel-10K Dataset

The Marvel-10K dataset consists of 501 stereo video sequences from 5 Marvel movies: *Ant-Man and the Wasp: Quantumania* (85 scenes), *Black Panther: Wakanda Forever* (83 scenes), *Doctor Strange in the Multiverse of Madness* (119 scenes), *Guardians of the Galaxy Vol. 3* (101 scenes), and *Thor: Love and Thunder* (113 scenes). Since movie frames are highly correlated within shots, we subsample them to select meaningful frames and exclude the studio intros, credits, and black frames. Each video sequence corresponds to a single shot and consists of 25 stereo pairs. For stereo conversion evaluation, we use left-view images as inputs and the right-view images as ground truth. As shown in Fig. 1, the Marvel-10K dataset features computer-generated characters, intense motions, complex lighting, and uncommon cinematic scenes, making it highly challenging and suitable for evaluating algorithms in real-world applications such as film production.

Figure 1. **Example images in the Marvel-10K dataset.**



(a) Visualization of depth label generation

(b) Comparisons of flying pixels in point renders

Figure 2. **Depth label generation.** We design alpha masking to produce sharp boundary labels for depth refinement, producing sharp depth edges and significantly reducing flying pixels in depth-fixed results.

1.2. Depth Label Generation

Our goal is to generate sharp depth labels for soft boundaries. We use pure foreground colors (unpremultiplied, *i.e.*, not mixed with background colors) and fill the background ($\alpha = 0$) with green to produce high-contrast and dilated boundaries (Fig. 2a). Although depth models show suboptimal performance with intermediate values at boundaries

(original d_{FG}), we multiply d_{FG} by the alpha mask M_α to filter unreliable boundary depth (Eq. (2) in the main paper). This produces sharp depth edges (masked d_{FG}) and significantly reduces flying pixels in depth-fixed results (Fig. 2b).

1.3. Color Fuser

We provide more implementation details about the dual skip module in our color fuser. Given the inpainted image $I_{inpaint}$ and the warped image I_{warp}, we first extract multi-scale features $\{F_{inpaint,i}\}_i, \{F_{warp,i}\}_i$ using the frozen VAE encoder. We also generate multi-scale warped masks $\{M_{warp,i}\}_i$ by resizing the original mask to each feature scale via nearest neighbor downsampling. Finally, we use an additional residual block in the VAE decoder to fuse the skipped features and masks at each feature scale, and inject the fused feature into the VAE decoder in a residual fashion,

$$F = F_{dec} + \text{ResBlock}(F_{dec}, F_{inpaint}, F_{warp}, M_{warp}).$$

$\text{ResBlock}(\cdot)$ indicates a residual block. F_{dec} and F correspond to the original decoder feature and the fused feature, respectively. Zero initialization is applied for the additional residual blocks during training.

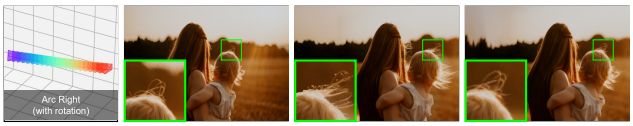
2. More Analysis on Model Performance

2.1. Performance under Large Viewpoint Changes

Following ReCamMaster [1], we evaluate novel view synthesis across 10 different camera trajectories featuring large viewpoint changes. Tab. 1 and Fig. 3 demonstrate our state-of-the-art performance and robustness in preserving soft boundary details.

Table 1. **Novel view synthesis performance under large viewpoint changes**. We employ 10 camera trajectories in the ReCamMaster evaluation protocol [1]. Metrics are FID ↓ / CLIP-F ↑. The best results are marked.

Dataset	ReCamMaster	SplatDiff	Ours
AIM-500	82.88 / 98.02	34.86 / 98.94	32.49 / 99.25
P3M-10K	116.80 / 97.18	57.57 / 98.61	50.33 / 98.85

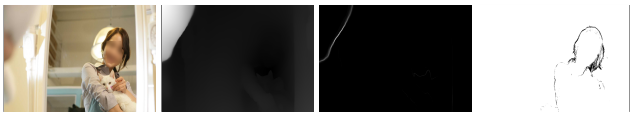


Trajectory Input SplatDiff Ours

Figure 3. **Visual results under large viewpoint changes.**

2.2. Soft Boundary Detection in Regions with Low Depth Gradients

Depth fixer leverages both geometry and image semantics to detect problematic (instead of all) soft boundaries for fixing.



Input Image Input Depth Sobel Edge Estimated Gate G

Figure 4. **Soft boundaries in low-gradient depth regions**. Human faces are manually blurred to protect privacy.

The Sobel edges of depth maps serve as guidance rather than a hard constraint. Fig. 4 shows that HairGuard can detect soft boundaries (gate $G < 1$) even in regions with negligible depth gradients.

2.3. Robustness and Generalization

Since our depth fixer is trained on a relatively small synthetic dataset (approximately 20K samples), one concern is its robustness and generalization ability in complex scenes. To this end, we evaluate the performance of the depth fixer on the challenging Marvel-10K dataset, which is not seen during training. As shown in Fig. 5, the depth fixer can automatically identify soft boundary regions in various scenarios. In scenes without soft boundaries (*e.g.*, the top two rows in Fig. 5), the depth fixer maintains the depth quality of the base depth model for robust zero-shot estimation. In complex scenes such as bright/dark environments, occlusions, and multiple targets (bottom three rows in Fig. 5), our depth fixer still exhibits promising performance in extracting and fixing soft boundaries, showcasing its robustness in real-world applications. This is attributed to the decoupling of depth estimation and soft-boundary refinement in our depth fixer. Thanks to the proposed gated residual mechanism, we can leverage the base depth model for zero-shot transfer and focus solely on refining soft boundaries, thereby achieving strong generalization performance with efficient training.

Table 2. **Plug-and-play performance of depth fixer on MoGe-2 [8] and PPD [9]**. Metrics are DBE_comp ↓ / DBE_acc ↓ / EP ↑ / ER ↑. Best results are marked.

Method	AIM-500	P3M-10K
MoGe-2	7.38 / 2.89 / 26.27 / 10.49	6.82 / 2.33 / 31.94 / 14.64
+ Depth Fixer	6.92 / 2.08 / 38.10 / 14.87	6.73 / 1.76 / 41.82 / 16.40
PPD	6.97 / 3.24 / 23.77 / 12.86	6.66 / 2.52 / 30.71 / 17.20
+ Depth Fixer	5.29 / 1.48 / 48.69 / 26.63	5.09 / 1.23 / 52.51 / 29.91

2.4. Plug-and-Play Performance

Benefiting from the gated residual module, our depth fixer can be applied to improve depth predictions from different depth models in a plug-and-play manner. As visualized in Fig. 6a, we apply the depth fixer to two depth models with different characteristics: Depth Pro captures better de-

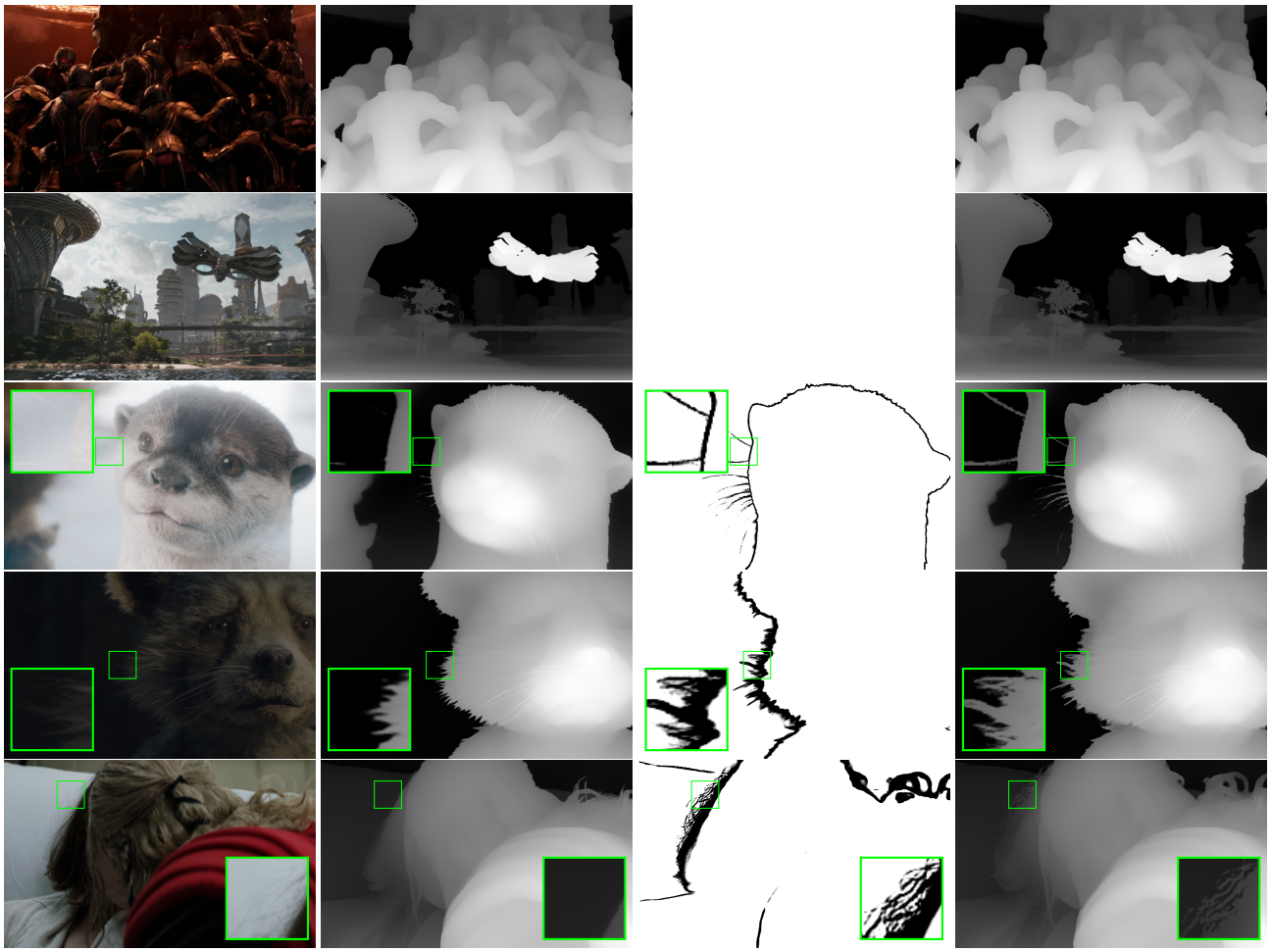


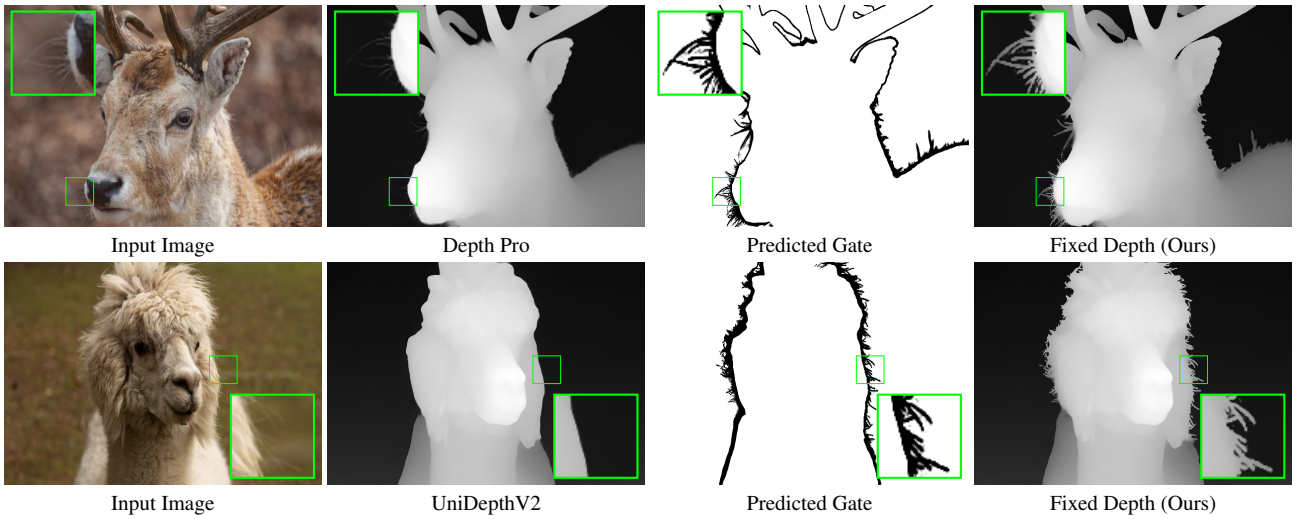
Figure 5. **Performance of depth fixer under challenging scenarios.** The regions with the predicted gate $G < 1$ indicate the estimated soft boundary regions. Even under complex environments, *e.g.*, heavy occlusion, extreme lighting conditions, and multiple targets, our depth fixer can automatically identify soft boundary regions and perform precise fixing.

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

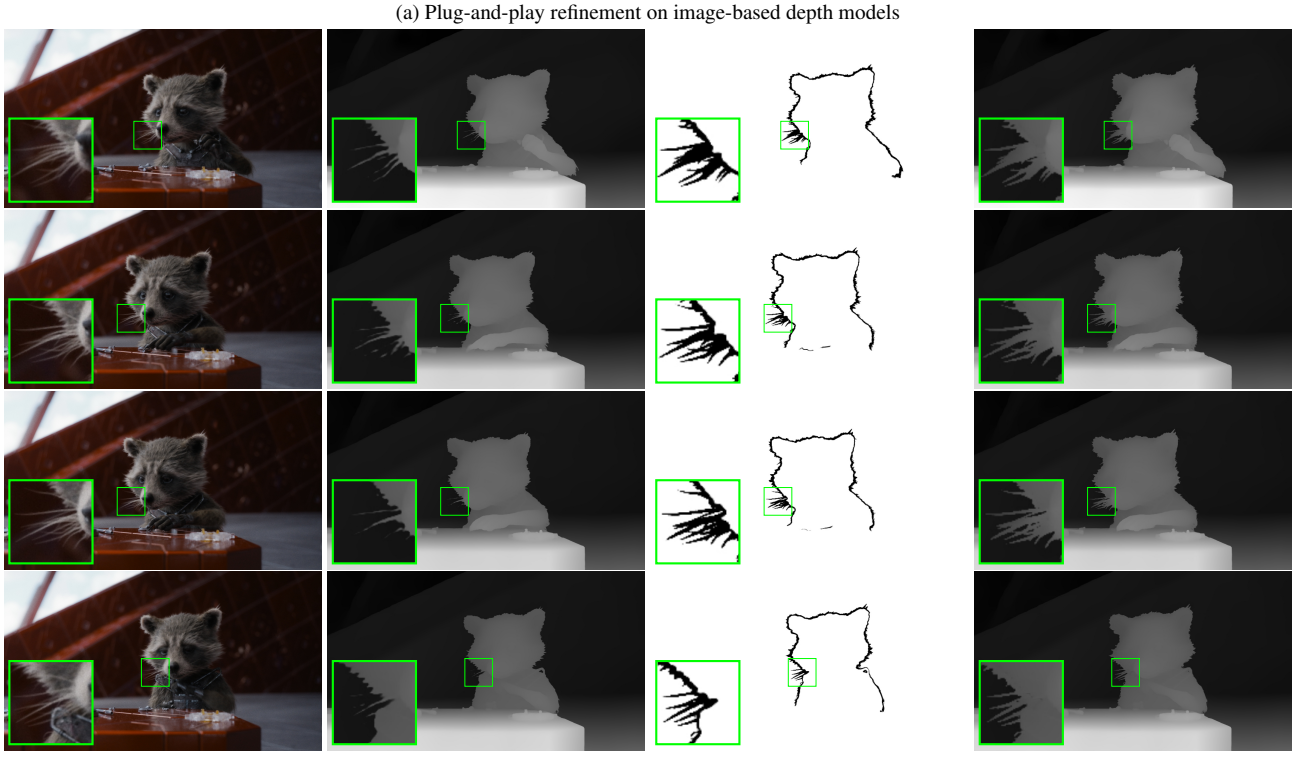
Method	Stereo Image Conversion						Stereo Video Conversion					
	PSNR ↑	SSIM ↑	RMSE ↓	LPIPS ↓	DISTS ↓	SIoU ↑	PSNR ↑	SSIM ↑	RMSE ↓	LPIPS ↓	DISTS ↓	SIoU ↑
SplatDiff [14]	36.23	0.8857	4.06	0.1116	0.0435	0.3259	36.24	0.8858	4.06	0.1114	0.0437	0.3280
SplatDiff+Depth Fixer (Ours)	36.38	0.8915	4.00	0.0974	0.0348	0.3309	36.39	0.8917	3.99	0.0972	0.0351	0.3326

tails but often predicts inaccurate depth values in boundaries [2], and UniDepthV2 tends to produce depth results with smoothed boundaries [7]. We also test the plug-and-play performance of the depth fixer on the very recent depth methods: MoGe-2 [8] and Pixel-Perfect Depth (PPD) [9]. Tab. 2 demonstrates our state-of-the-art performance on MoGe-2 and PPD (settings match Tab. 1 in the main paper). Despite the different distributions of depth maps, our depth fixer maintains robust performance in predicting soft boundary regions and fixing depth details.

We further evaluate the plug-and-play capability of the depth fixer on video depth models, *e.g.*, Video Depth Anything [3]. Although the depth fixer is trained only on image datasets, it still exhibits remarkable performance in improving video depth results, as shown in Fig. 6b. Thanks to the gated residual mechanism, our depth fixer only corrects the depth in soft boundary regions while preserving the temporal consistency of video depth results. Besides, the depth fixer shows stable performance in estimating soft boundary regions even under occluded scenes (*e.g.*, see the predicted gate maps in Fig. 6b), demonstrating its robustness in complex scenarios.



(a) Plug-and-play refinement on image-based depth models



(b) Plug-and-play refinement on video-based depth models

Figure 6. **Plug-and-play performance of depth fixer.** The depth fixer can be integrated with different depth models, *e.g.*, image-based models in (a) and video-based models in (b), in a plug-and-play fashion for soft boundary refinement. Although the depth fixer is trained only on image datasets, it can be directly applied to improve video-based models such as Video Depth Anything [3], without additional re-training. Leveraging the gated residual mechanism, the depth fixer preserves the temporal consistency of the video depth model while achieving stable performance in identifying soft boundaries and recovering fine-grained details, even in complex scenes with occlusions.

In addition to enhancing depth estimation methods, our depth fixer can also be integrated with novel view synthesis models for performance improvement. For instance, we combine depth fixer with the previous novel view synthesis approach SplatDiff [14], and evaluate its performance on the Marvel-10K dataset. Since the depth fixer improves the warping results by fixing soft boundary details in depth (detailed in Sec. 3.1), its combination with SplatDiff shows a

Table 4. **Complexity comparison** with previous state-of-the-art novel view synthesis methods on the Marvel-10K dataset at a resolution of 384×640, evaluated using an NVIDIA GeForce RTX 4090 GPU. For diffusion-based methods, we only take into account the model sizes of the latent diffusion model and the VAE model. * means that the method runs out of memory, and thus we perform inference at a lower resolution 256×448 for reference. The complexity of each component in HairGuard *i.e.*, depth fixer, scene painter, and color fuser, is also reported. Since the three components are applied sequentially, the peak GPU memory of HairGuard equals that of the scene painter. The best and second-best results are marked.

Method	Model Size	Peak GPU Mem.	Infer. Speed
ViewCrafter [13]	2.22 B	14.91 G	47.83 s
NVS-Solver* [12]	2.25 B	21.60 G	100.00 s
ReCamMaster [1]	1.51 B	17.11 G	684.44 s
SplatDiff [14]	2.28 B	22.88 G	52.28 s
HairGuard (Ours)	1.86 B	10.65 G	95.23 s
Depth Fixer (Ours)	0.32 B	1.84 G	0.03 s
Scene Painter (Ours)	1.44 B	10.65 G	95.19 s
Color Fuser (Ours)	0.10 B	3.15 G	0.01 s

Table 5. **Warping performance** using different depth maps on the Marvel-10K dataset. Metrics are computed only on the soft boundary regions. Best results are marked.

Method	PSNR ↑	SSIM ↑	RMSE ↓
Depth Anything V2 [10]	30.18	0.4495	8.08
Depth Anything V2+Depth Fixer (Ours)	31.07	0.5140	7.36
Depth Pro [2]	31.12	0.5591	7.28
Depth Pro+Depth Fixer (Ours)	31.77	0.6144	6.79
UniDepthV2 [7]	31.31	0.5637	7.12
UniDepthV2+Depth Fixer (Ours)	32.23	0.6261	6.46

consistent performance gain across all metrics, as reported in Tab. 3.

2.5. Computational Complexity

In Tab. 4, we compare the computational complexity, *i.e.*, model size, peak GPU memory, and inference speed, of HairGuard with previous state-of-the-art novel view synthesis methods. We further break down the complexity of each component in HairGuard, and the results show that the scene painter dominates the computational cost in our framework. Since we apply the depth fixer, scene painter, and color fuser in a sequential way, the peak GPU memory of HairGuard equals that of the scene painter. As our primary contributions lie in the depth fixer and the color fuser, the scene painter can be replaced with a more lightweight variant for better efficiency. In summary, our HairGuard achieves state-of-the-art performance while maintaining competitive computational efficiency, as demonstrated in Tab. 4.

Table 6. **Ablation study of depth fixer** on the Marvel-10K dataset. The ablations about gated residual, loss function, model prior, edge guidance, and alpha threshold correspond to experiments #1-3, #4-5, #6-7, #8-9, and #10-12, respectively. We use Depth Anything V2 as the base depth model [10]. Metrics are computed only on the soft boundary regions. Best results are marked.

Exp	Strategies	Marvel-10K		
		PSNR ↑	SSIM ↑	RMSE ↓
#1	Direct Prediction	30.66	0.5124	7.67
#2	Vanilla Residual	30.37	0.5009	7.88
#3	Gated Residual (Ours)	31.07	0.5140	7.36
#4	$\mathcal{L}_1$ Only	30.58	0.5057	7.74
#5	$\mathcal{L}_1 + \mathcal{L}_\alpha$ (Ours)	31.07	0.5140	7.36
#6	w/o Model Prior	30.26	0.4668	8.00
#7	w/ Model Prior (Ours)	31.07	0.5140	7.36
#8	w/o Edge Guidance	30.66	0.5057	7.67
#9	w/ Edge Guidance (Ours)	31.07	0.5140	7.36
#10	$\alpha_{th} = 0.1$	30.43	0.4774	7.87
#11	$\alpha_{th} = 0.05$	30.55	0.4917	7.76
#12	$\alpha_{th} = 0.02$ (Ours)	31.07	0.5140	7.36

3. More Experiments on Depth Fixer

3.1. Warping Performance

In this section, we analyze the influence of the depth fixer on view synthesis tasks. To focus on the impact of depth maps, we directly assess the quality of the warped images, without applying the scene painter and color fuser. In addition, since the proposed depth fixer only modifies the depth on the predicted soft boundary regions, *i.e.*, regions with gate $G < 1$, we compute pixel-level metrics only on these regions. We apply our depth fixer in a plug-and-play fashion to improve the prediction from three state-of-the-art depth models (Depth Anything V2 [10], Depth Pro [2], and UniDepthV2 [7]), and compare the forward warping performance on the Marvel-10K dataset. As shown in Tab. 5, our depth fixer helps preserve more soft boundary details during forward warping, leading to consistent and significant improvements across different base depth models.

3.2. Gated Residual

Following the same experimental setting in Sec. 3.1, we compare the performance of the depth fixer with different output mechanisms. Although the direct prediction and vanilla residual mechanisms help improve the depth on the soft boundary regions, they often cover redundant background regions and fail to capture the fine-grained details, as illustrated in Fig. 4c of the main paper. By accurately localizing the soft boundary regions with the estimated gate map, our gated residual facilitates precise depth refinement and achieves the best performance as shown in Tab. 6 (#3 *vs*. #1-2).

3.3. Loss Function

We train the depth fixer with the ℓ_1 loss $\mathcal{L}_1$ and the image matting loss $\mathcal{L}_\alpha$ [11]. Specifically, the image matting loss $\mathcal{L}_\alpha$ is formulated as

$$\mathcal{L}_\alpha = \mathcal{L}_1 + \mathcal{L}_{lap} + \mathcal{L}_{gp}, \tag{1}$$

where $\mathcal{L}_{lap}, \mathcal{L}_{gp}$ indicate the Laplacian loss [6] and the gradient loss [4], respectively. Inspired by the success of such a loss combination in the image matting task [11], we adopt it to improve the detail extraction performance of our depth fixer. To verify its effectiveness, we train an additional model with $\mathcal{L}_1$ loss only and keep the other training settings unchanged. The results in Tab. 6 show the better performance of the proposed loss combination (#5 *vs*. #4).

3.4. Model Prior

Identifying soft boundary regions is a challenging task, relying on a comprehensive understanding of semantic context and geometric layout. To this end, we initialize the feature branch of our depth fixer with the pre-trained Depth Anything V2 [10], which has been trained on large-scale datasets to acquire robust image and geometry priors. Benefiting from this, our depth fixer achieves strong performance with efficient training (only ∼20K training samples are used), as shown in Tab. 6 (#7 *vs*. #6).

3.5. Edge Guidance

Object boundaries, especially regions with significant depth variations, play a crucial role in 3D tasks like view synthesis, where disocclusions and geometric distortions commonly occur. Thus, we extract edge cues from the input depth to guide the depth fixer. The depth gradients provided by the edge guidance enable more accurate localization of soft boundaries, leading to improved warping performance (#9 *vs*. #8 in Tab. 6).

3.6. Alpha Threshold

The alpha threshold α_{th} used in generating ground-truth depth is critical to the performance of depth fixer. Unlike hard boundaries, where pixels are discretely assigned to either foreground or background, pixels at soft boundaries exhibit a blend of both (see Eq. (1) in the main paper), effectively corresponding to two depth layers. Although regions with low (but non-zero) alpha values may visually resemble the background (*e.g*., red lines in Fig. 7), we prioritize treating them as foreground by using a low α_{th}. As shown in Fig. 8, the model trained with a higher α_{th} exhibit finer delineation of depth boundaries with less redundant background regions, but it tends to ignore the areas with low opacity, *e.g*., very thin hair. In our design, we opt for a lower α_{th} to preserve as many soft boundary details as possible, which shows the best warping performance in Tab. 6 (#12 *vs*. #10-11). The scene painter and the color fuser are then employed to fix redundant background regions during view synthesis, as illustrated in Fig. 5c of the main paper. Nevertheless, a higher α_{th} can be used to train the depth fixer for different tasks, *e.g*., 3D segmentation or point cloud reconstruction, where precise depth boundaries are preferred.

Figure 7. **Soft-Boundary pixels belong to two depth layers.** Unlike hard boundaries, where pixels are discretely assigned to either foreground or background, pixels at soft boundaries exhibit a blend of both and thus correspond to two depth layers.

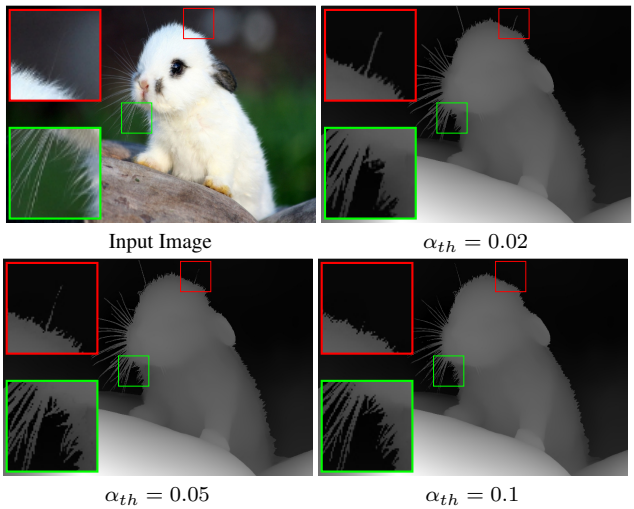
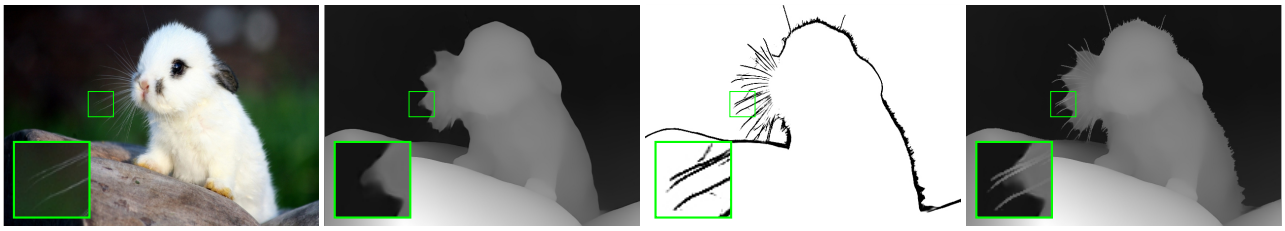


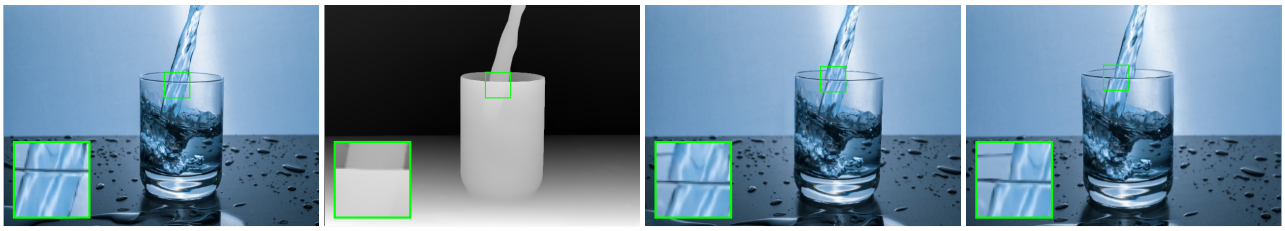
Figure 8. **Performance of the depth fixer trained with different alpha thresholds.** A higher threshold α_{th} leads to less redundant background in the depth map (*e.g*., green box), while a lower threshold α_{th} improves the coverage of fine-grained details, *e.g*., the very thin hair in the red box.

Table 7. **Ablation study of color fuser** on the Marvel-10K dataset. The ablations about VAE prior and skip mechanisms correspond to experiments #1-2 and #3-5, respectively. Best results are marked.

Exp	Strategies	Marvel-10K			
		PSNR ↑	LPIPS ↓	DISTS ↓	FID ↓
#1	w/o VAE Prior	36.61	0.0965	0.0366	7.99
#2	w/ VAE Prior (Ours)	36.59	0.0909	0.0331	7.19
#3	w/o Skip	34.96	0.1664	0.0807	18.11
#4	Sinlge Skip	36.46	0.0919	0.0337	7.34
#5	Dual Skip (Ours)	36.59	0.0909	0.0331	7.19



(a) Depth errors beyond soft boundaries



(b) Single-layer depth representation

Figure 9. **Failure cases.** (a) Since the depth fixer only fixes the depth in the soft boundaries (represented by the regions with gate $G < 1$), it is difficult to correct depth errors beyond soft boundaries in the prediction of the base depth model. (b) Due to the limitation of single-layer depth representation, the synthesized novel view might fail to correct the geometric errors caused by forward warping.

4. More Experiments on Color Fuser

4.1. VAE Prior

We build the color fuser upon a pre-trained VAE to harness its reconstruction prior for better view synthesis performance. To investigate the impact of the VAE prior, we train an additional color fuser from scratch using the same training settings. As shown in Tab. 7, the color fuser with VAE prior significantly outperforms its counterpart in visual quality (#2 *vs*. #1).

4.2. Dual Skip

Based on the VAE architecture, we further design a dual skip module to utilize the fine-grained features of the inpainted and warped images. To verify its effectiveness, we train two additional variants: one without skip connections (#3 in Tab. 7) and one with a single skip connection (#4). For model #3, we expand the input channel of the VAE encoder and concatenate the inpainted image, warped image, and warped mask as its input. Regarding model #4, we add a single skip to utilize the multi-scale features of the warped images. The results in Tab. 7 show that model #4 achieves a significant performance gain over model #3 by alleviating detail compression in the VAE encoding. By further exploiting the features of inpainted images, our dual skip module yields the best reconstruction performance with high-quality texture details.

5. Limitation and Discussion

Despite the remarkable performance achieved by HairGuard, some limitations remain:

- *Depth errors beyond soft boundaries:* The depth fixer relies on the gated residual mechanism to locate and fix soft boundary details, which benefits precise refinement and plug-and-play deployment. However, it is difficult for the depth fixer to correct depth errors beyond the soft boundary regions, as depicted in Fig. 9a. A possible solution is to train a depth fixer specialized for a given depth model, *e.g*., by using the model's predictions instead of the synthesized inputs during training. Thus, the depth fixer could better adapt to the characteristics of the base depth model and achieve better fixing performance.
- *Single-layer depth representation:* For view synthesis, we propose a color fuser to utilize the fine-grained texture from the warped images. However, due to the single-layer depth representation, the naive forward warping approach may produce geometric distortions in complex scenes containing multiple depth layers per pixel, *e.g*., transparent objects as shown in Fig. 9b. We attempted to address this limitation by estimating layered outputs comprising foreground color and depth, background color and depth, and an opacity map for composition. While this layered representation demonstrated advantages in certain cases, our trial experiments showed that it suffers from limited generalization capability, likely due to the increased complexity of the estimation. Thus, a potential solution is to collect large-scale training datasets to gain a strong prior

(a) Visual results on the AIM-500 dataset

(b) Visual results on the Marvel-10K dataset

Figure 10. **Visual results of ablation study** on the AIM-500 and Marvel-10K datasets. Due to depth estimation errors, the original warped images often contain broken or distorted structures in thin hairs. Our depth fixer improves the warping performance by fixing the soft boundary regions in the depth. The scene painter is employed to fill disoccluded regions in the warped images, but the inpainted results often suffer from hallucinated details that are inconsistent with the input image (*e.g.*, see hairs in the green box, particularly in the Marvel-10K examples). By adaptively combining the warped and inpainted images, the color fuser produces high-quality results with consistent texture and geometry.

for robust performance. Another possible direction is to employ dense 3D representations, *e.g.*, 3D Gaussians [5], to handle occlusions and overlapping surfaces.

6. More Visual Results

6.1. Ablation Study

Fig. 10 provides visual results for the ablation study conducted in the main paper (detailed in Tab. 4 of the main paper). Since depth quality is critical for forward warping performance, depth estimation errors in Depth Anything V2 [10] often result in distorted structures in the soft boundary regions like thin hairs. By fixing depth details via the proposed depth fixer, better hair structures are preserved in the warped images, as shown in the green box of Fig. 10. The scene painter is then applied to generate realistic contents for the disoccluded regions. However, the inpainted images often exhibit different texture details due to diffusion hallucination and pixel-to-latent compression. To this end, we propose a color fuser that adaptively combines the warped and inpainted images, generating novel views with consistent geometry and high-fidelity textures.

6.2. Monocular Depth Estimation

We provide more visual results of monocular depth estimation on the AIM-500 and P3M-10K datasets in Fig. 11 and Fig. 12, respectively. Compared with previous methods, our depth fixer shows robust performance in capturing soft boundary details across diverse targets and scenes. In some challenging cases with very thin hair structures, *e.g.*, top few rows of Fig. 12, the depth fixer still recovers fine-grained depth details with sharp boundaries.

6.3. Stereo Conversion

Fig. 13 compares the stereo conversion performance of HairGuard with the state-of-the-art methods on the Marvel-10K dataset. Due to the generative nature of the underlying models, previous stereo conversion approaches often suffer from texture hallucination and degraded details in the conversion results, *e.g.*, see the top two rows in Fig. 13. By

utilizing the fine-grained details of warped images via the color fuser, our HairGuard achieves high-quality stereo conversion performance with consistent geometry and texture.

6.4. Novel View Synthesis

We show more qualitative comparisons of novel view synthesis on the challenging AIM-500 and P3M-10K datasets in Fig. 14 and Fig. 15. Previous approaches often produce hallucinated textures that are inconsistent with the input image, *e.g.*, see ViewCrafter [13] and ReCamMaster [1] in the top few rows of Fig. 14. Although the recent method SplatDiff recovers better details [14], its performance is highly dependent on the quality of the estimated depth maps. Hence, the depth errors in soft boundary regions often lead to artifacts in the synthesized novel views, *e.g.*, top few rows in Fig. 15. In contrast, the proposed HairGuard first fixes depth in the soft boundary regions to ensure geometry consistency, and then utilizes the color fuser to recover high-fidelity texture details, achieving state-of-the-art novel view synthesis performance.

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, 5, 9

[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, 5

[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. 3, 4

[4] Yutong Dai, Brian Price, He Zhang, and Chunhua Shen. Boosting robustness of image matting with context assembling and strong data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11707–11716, 2022. 6

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 8

[6] 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. 6

[7] 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. 3, 5

[8] 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 *NIPS*, 2025. 2, 3

[9] 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 *NIPS*, 2025. 2, 3

[10] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 5, 6, 8

[11] Jingfeng Yao, Xinggang Wang, Shusheng Yang, and Baoyuan Wang. Vitmatte: Boosting image matting with pretrained plain vision transformers. *Information Fusion*, 103: 102091, 2024. 6

[12] Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In *ICLR*, 2025. 5

[13] 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. 5, 9

[14] 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. 3, 4, 5, 9

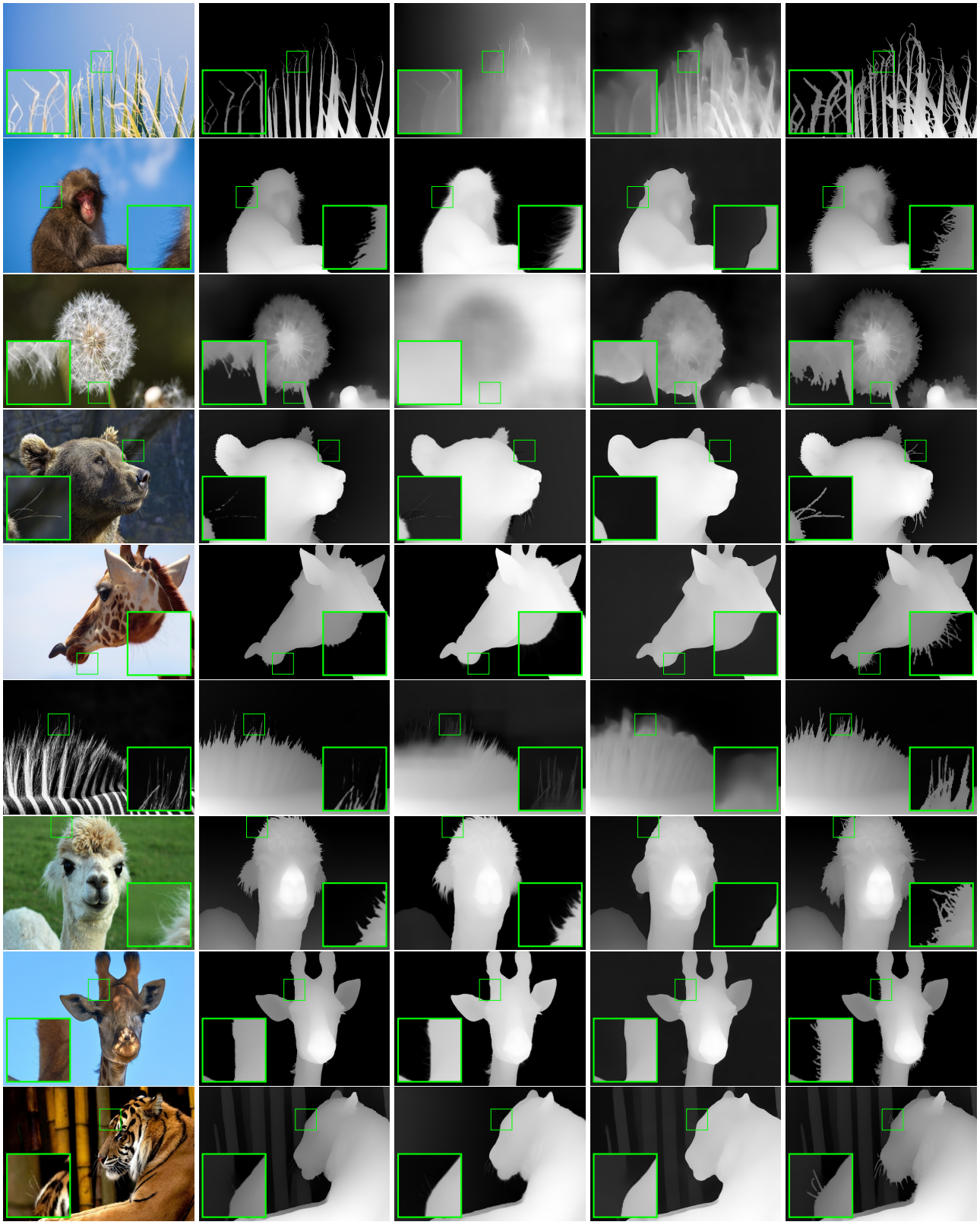


Figure 11. **Qualitative comparison of depth estimation** on the AIM-500 dataset.

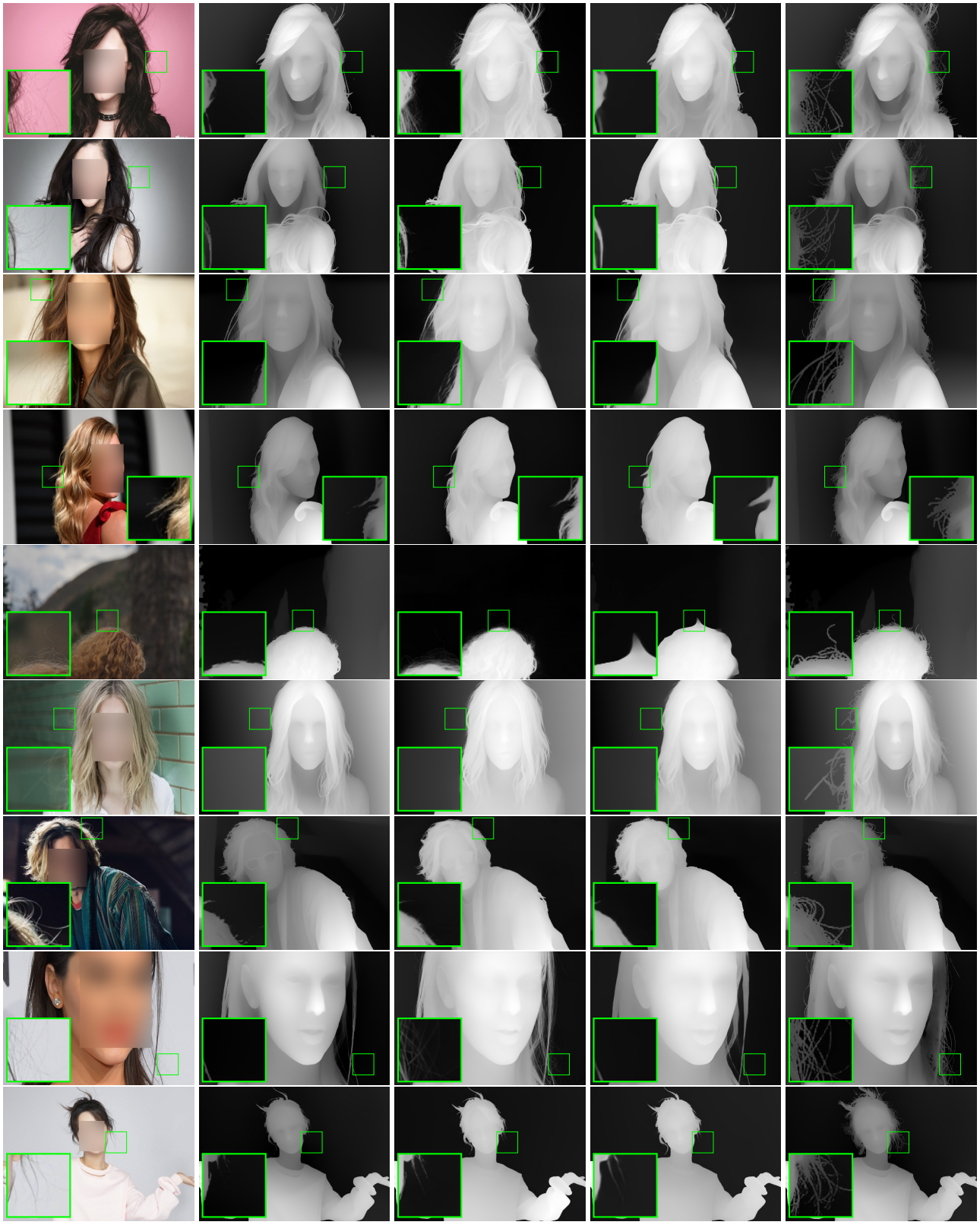


Figure 12. **Qualitative comparison of depth estimation** on the P3M-10K dataset. Human faces are manually blurred to protect privacy.

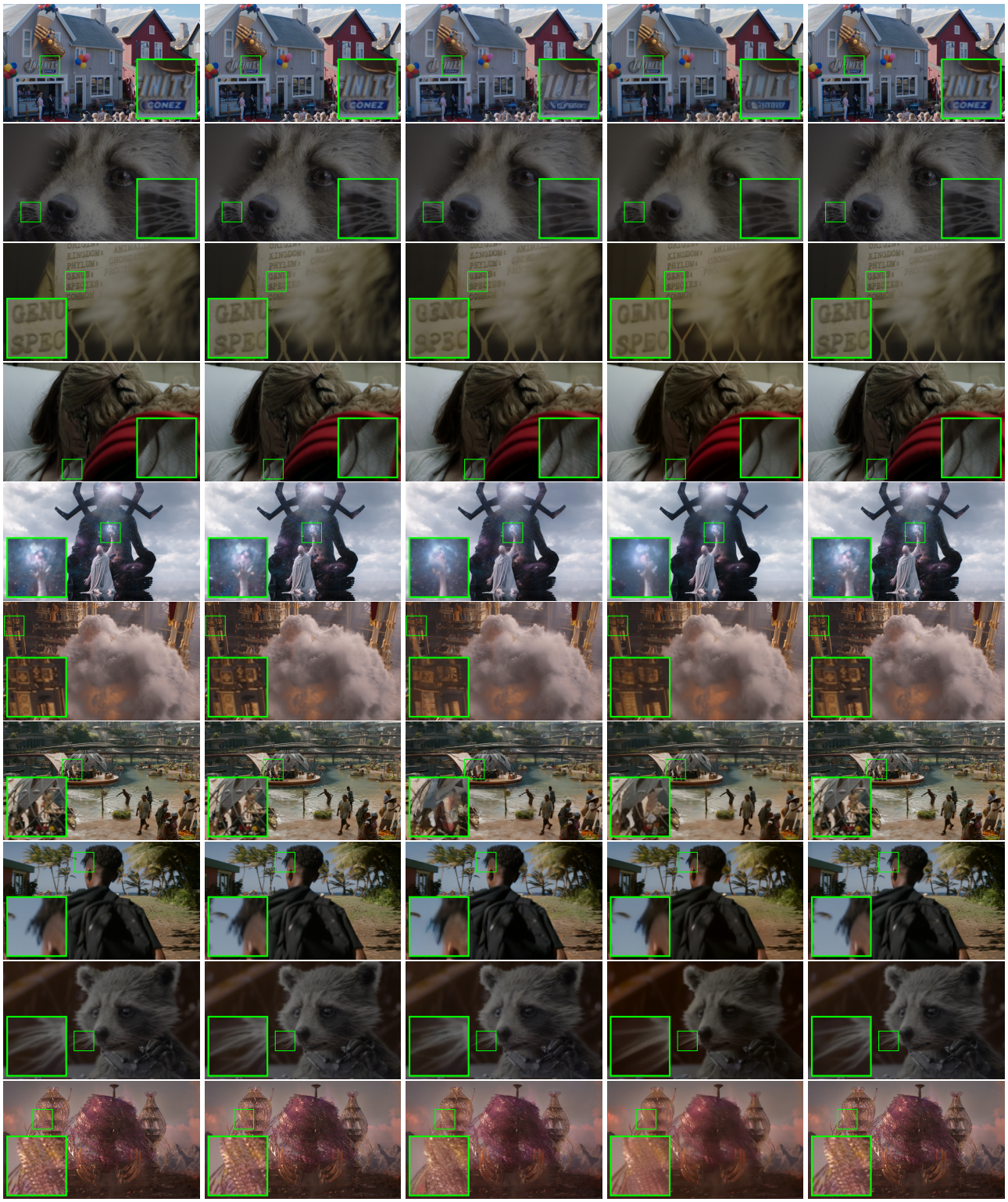


Figure 13. **Qualitative comparison of stereo conversion** on the Marvel-10K dataset.

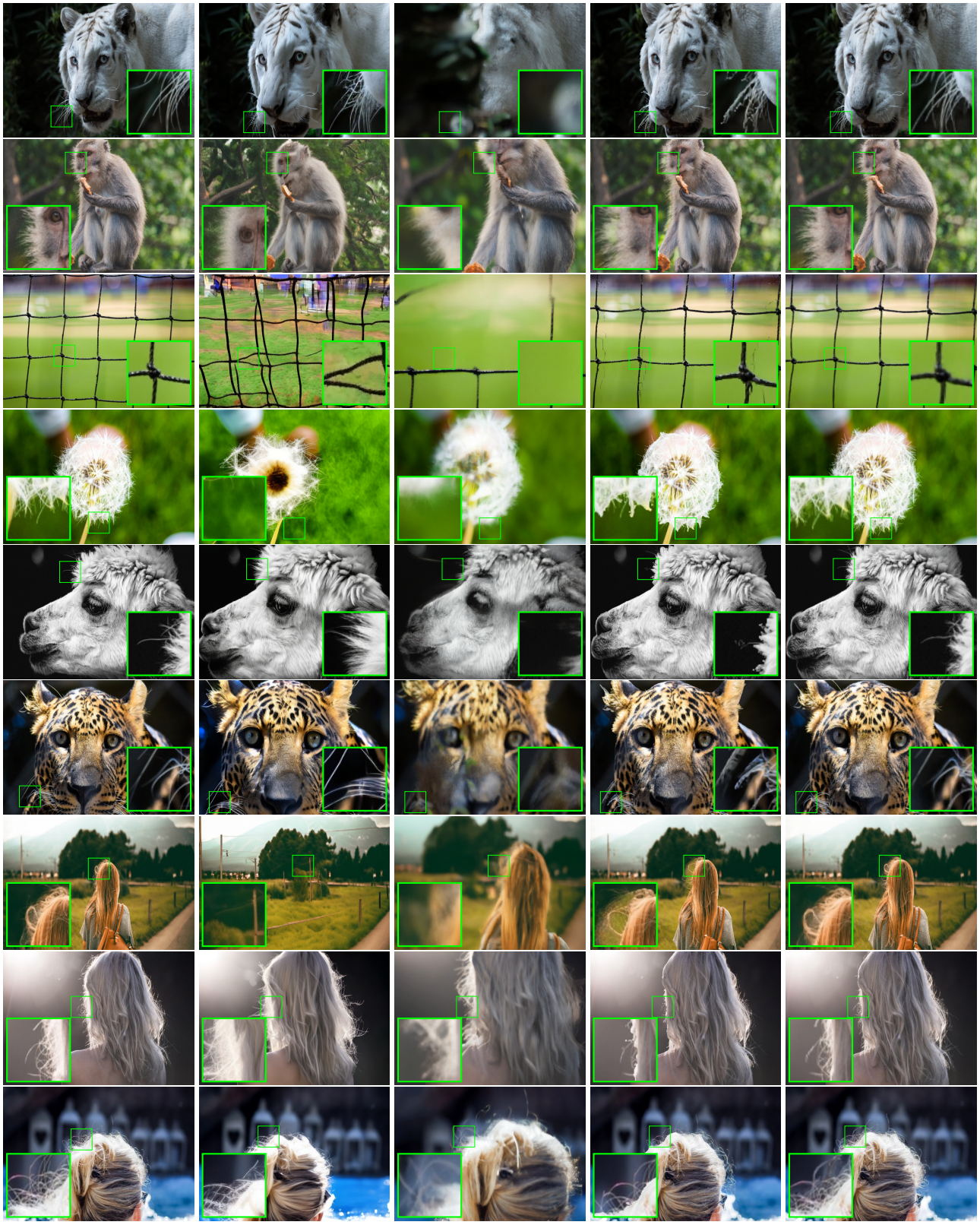


Figure 14. **Qualitative comparison of novel view synthesis** on the AIM-500 dataset.

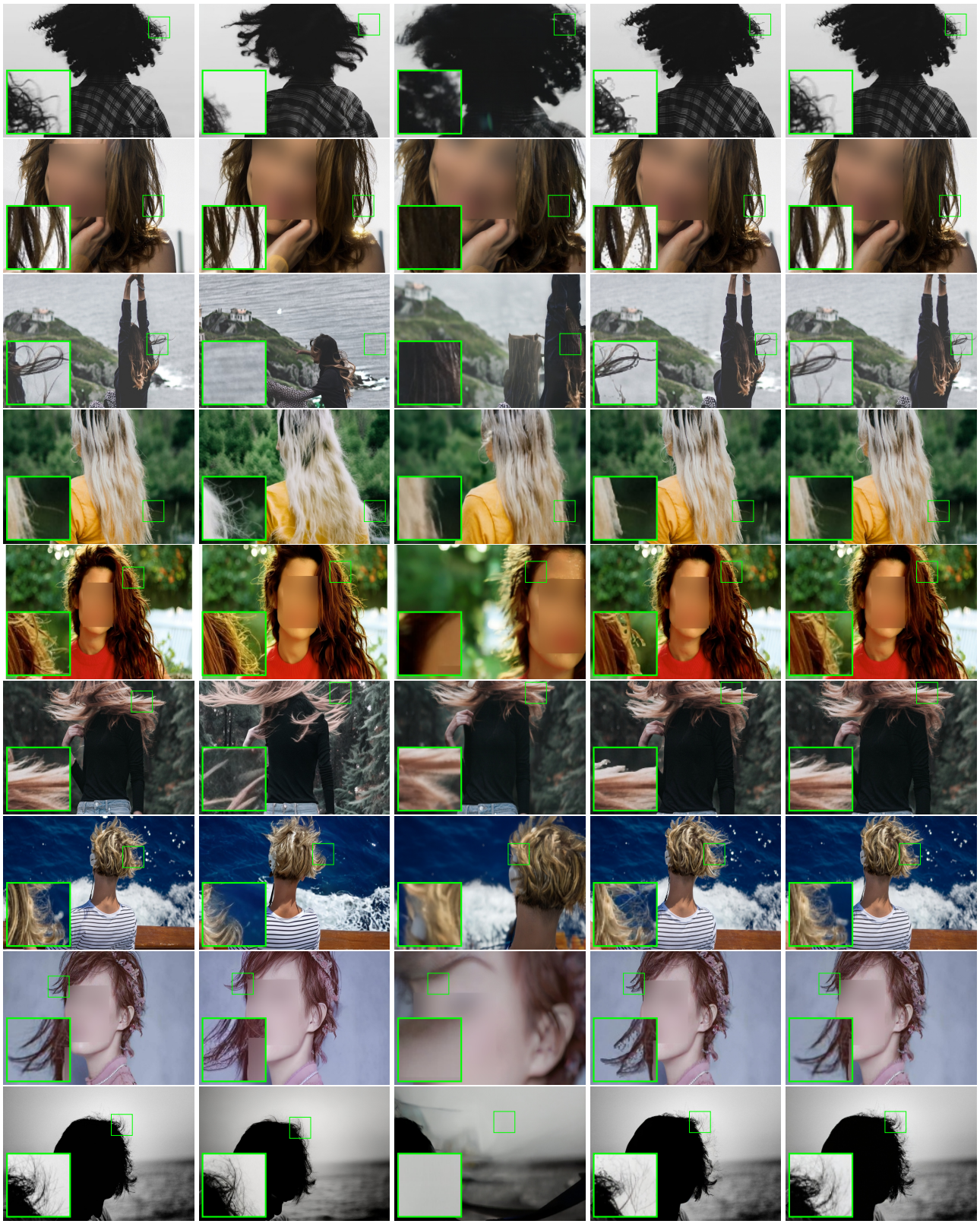


Figure 15. **Qualitative comparison of novel view synthesis** on the P3M-10K dataset. Human faces are manually blurred to protect privacy.