ZeroStereo: Zero-shot Stereo Matching from Single Images

Supplementary Material

Method	KITTI-15		Midd-T (H)		ETH3D	
	EPE	>3px	EPE	>2px	EPE	>1px
$\mathcal{L}_p$	1.03	4.77	0.90	4.95	0.25	2.09
$(1-\mathbf{C})\odot\mathcal{L}_p$	1.03	4.76	0.85	4.81	0.25	**2.08**
$(1-\mathbf{C})\odot\mathcal{L}_{np}$	**1.02**	**4.53**	**0.79**	**4.45**	**0.23**	2.13

Table 9. Analysis of ZeroStereo loss (trained with RAFT-Stereo [25]). We discuss the loss combinations based on $\mathcal{L}_d$.

6. Details of Image Synthesis

Image Resolution. The input resolution of Depth Anything V2 [63] and Stable Diffusion V2 Inpainting [39] is constrained, which may lead to object deformation when resizing images. To address this, we apply padding operations to adjust image dimensions while preserving their original aspect ratio. For example, when using Depth Anything V2, we pad images to ensure their height and width are divisible by 14. Additionally, high-resolution images, particularly those from the Mapillary Vistas [34], may exceed available GPU memory during inference. To mitigate this issue, we first downscale images proportionally to half or quarter of their original resolution, perform inference, and then upscale the outputs to restore the original dimensions.

Forward Warping. We utilize the source code of MfS-Stereo [54] to implement forward warping, including non-occlusion computation and depth sharpening. However, when applying a diffusion model for inpainting, we identify several challenges. First, despite advancements in monocular depth estimation, depth edges do not always align precisely with object boundaries. As a result, after forward warping, the edges of foreground objects may remain in their original positions. Second, the proximity between the inpainting mask and the warped foreground objects can mislead the diffusion model during inference. To mitigate these issues, we employ a simple yet effective approach: using the dilate function in OpenCV to inflate the pseudo-disparity map. This operation ensures that foreground objects and nearby background pixels move together during forward warping. Consequently, during inpainting, background pixels act as a buffer between the mask and the foreground, reducing misleading information. However, despite this refinement, the pre-trained diffusion model still produces ghosting artifacts and noise in many cases (Fig. 5). These artifacts can only be effectively addressed by fine-tuning the diffusion model.

7. Loss Analysis

In Sec. 3.5, we introduce the non-occlusion photometric loss $\mathcal{L}_{np}$ and the weighted final loss $\mathcal{L}_{Zero}$. However, their specific impact on stereo training has not been explicitly analyzed. As shown in Tab. 9, the methods listed from top to bottom correspond to: (1) applying the ordinary photometric loss $\mathcal{L}_p$, (2) using $\mathcal{L}_p$ with the weight $1-\mathbf{C}$, and (3) employing $\mathcal{L}_{np}$ with the weight $1-\mathbf{C}$.

Among these, $\mathcal{L}_p$ alone yields the worst performance across all datasets due to the absence of balanced weighting and its inability to handle ghost artifacts and inpainting pixels. Introducing the weight $1-\mathbf{C}$ mitigates these issues, leading to improved performance. The best results are achieved when masks are further applied to filter out ghost artifacts and inpainting pixels, highlighting the effectiveness of our proposed approach.

Method	AbsErr ↓	SSIM ↑	$\mathcal{L}_p$ ↓
StereoDiffusion [51]	0.082	0.269	0.323
Ours	**0.025**	**0.850**	**0.068**

Table 10. Reconstruction loss. We warp the synthesized right image with the pseudo disparity and compare it with the left image.

8. Discussion on Synthesis Methods

In this section, we discuss two synthesis methods: StereoDiffusion [51] and AdaMPI [14].

StereoDiffusion [51] is a training-free method that utilizes a pre-trained latent diffusion model to generate stereo pairs from a single image. It applies null-text inversion [32] for image editing, first reversing the diffusion process to obtain a latent representation of the input image and then applying forward diffusion to synthesize the right view. However, this approach has notable limitations. First, inference is computationally expensive. As shown in Tab. 4, synthesizing a 512×512 image takes approximately 30 seconds. Second, the null-text inversion process can unintentionally modify the left image, introducing content inconsistencies. As illustrated in Fig. 9, the original image lacks stones, yet both the generated left and right views erroneously include them. Similarly, fine details such as text often become distorted. Quantitative reconstruction loss measurements (Tab. 10) confirm these issues, showing significantly higher errors compared to our method. Moreover, using StereoDiffusion-generated stereo pairs for training stereo matching networks led to poor performance and convergence difficulties.

AdaMPI [14] generates multiplane images [69] (MPI) from a single input image from a single input image for novel view synthesis. However, as shown in Fig. 10, varying the camera motion ratios often introduces artifacts, particularly in occluded regions, where ghosting and trailing effects are prevalent. This suggests that the MPI approach

Method	Cloudy		Foggy		Rainy		Sunny	
	F	H	F	H	F	H	F	H
NS-RAFT-Stereo	8.81	2.95	18.18	3.41	**29.19**	**8.47**	7.42	**2.88**
Zero-RAFT-Stereo	**6.44**	**2.69**	**8.66**	**1.70**	30.10	11.71	**6.46**	3.15

Table 11. Zero-shot generalization performance on DrivingStereo under different weather. We utilize >3px All in comparisons.

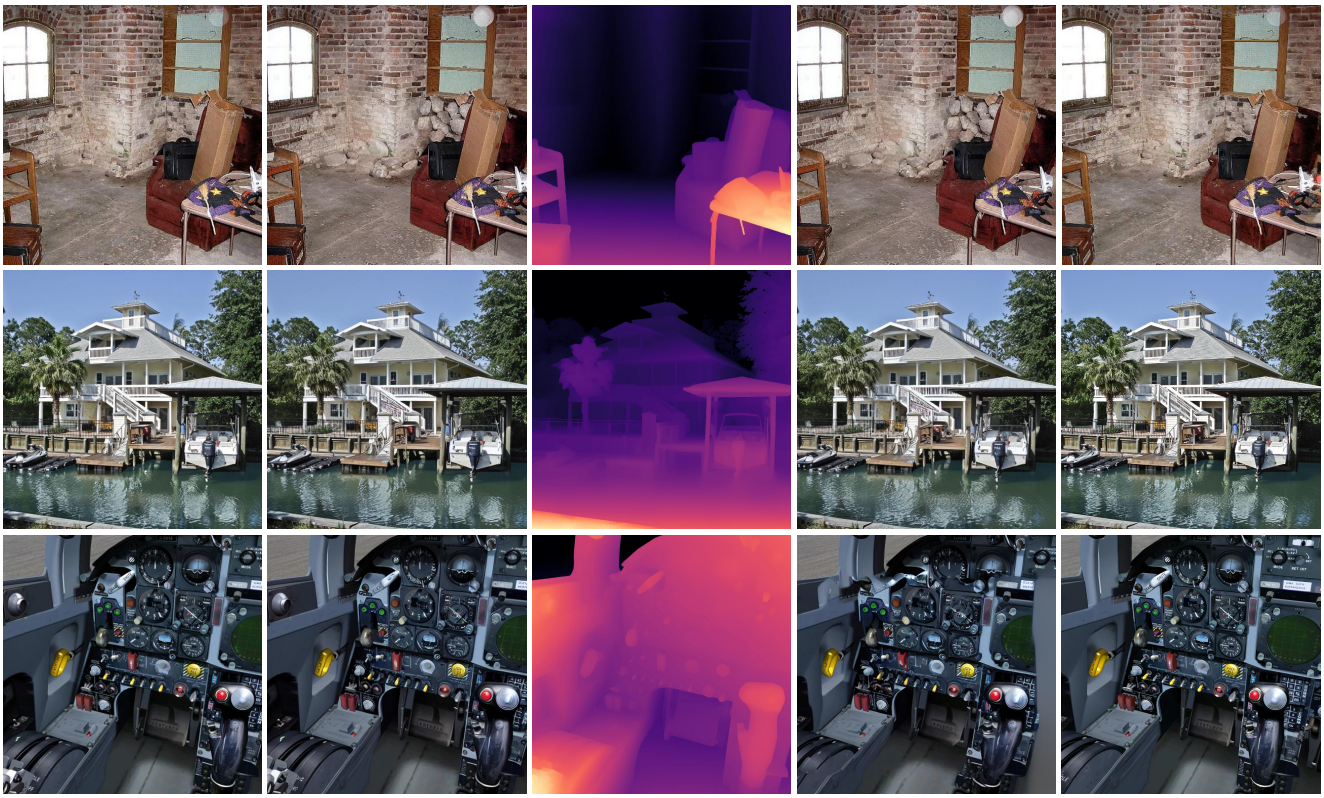


Figure 9. Visualization of StereoDiffusion [51].

struggles to reconstruct the scene's semantic structure accurately. As a result, MPI-based stereo generation is less suitable for training stereo matching models, as these artifacts compromise the quality and consistency needed for effective learning.

In summary, while StereoDiffusion [51] and AdaMPI [14] introduce innovative approaches for synthesizing stereo images from single inputs, both have significant limitations. StereoDiffusion suffers from high computational costs and content distortions, while AdaMPI struggles with semantic inconsistencies in occluded regions. These challenges highlight the need for more robust and accurate synthesis methods for stereo matching applications.

9. Additional Comparisons with NeRF-Stereo

In this section, we present additional comparisons with NeRF-Stereo [47], detailed Midd-T benchmark results, visualizations on KITTI and ETH3D, and zero-shot generalization performance on DrivingStereo [61].

For Midd-T, we report the performance of each sample in Tab. 12. Compared to NS-RAFT-Stereo [47], our Zero-RAFT-Stereo achieves improvements in nearly all cases. Notably, for samples where NS-RAFT-Stereo performs poorly, our method improves accuracy by almost 50%.

For KITTI and ETH3D, we provide visual comparisons between NS-RAFT-Stereo and Zero-RAFT-Stereo. As shown in Fig. 11, Fig. 12, Zero-RAFT-Stereo generates smoother and more accurate disparity maps with fewer artifacts and reduced noise. Notably, in the second row of Fig. 12, our model effectively removes the large disparity artifacts present in NS-RAFT-Stereo, particularly in the central dark region, demonstrating its superior handling of challenging textures and illumination variations.

Additionally, we evaluate both models on the DrivingStereo dataset under different weather conditions. As

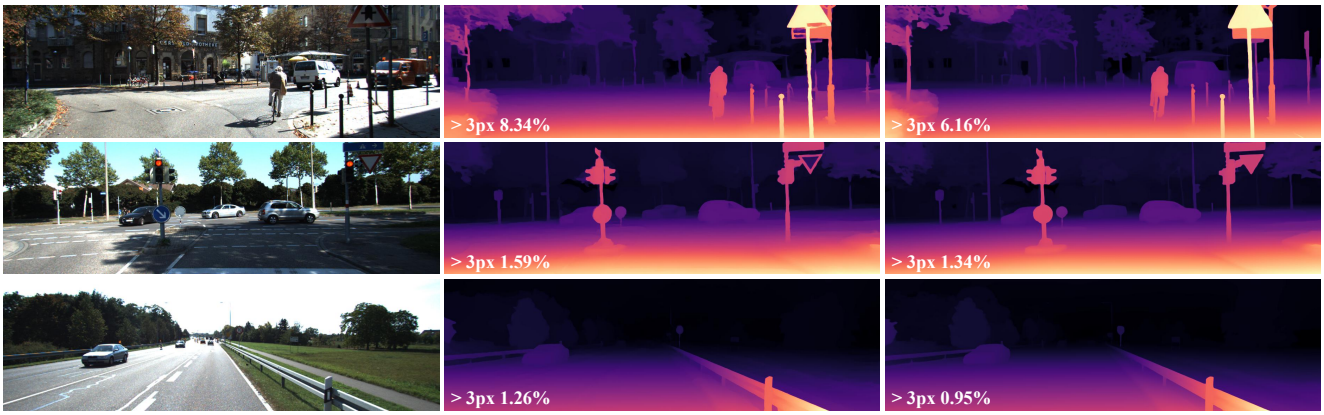
Method	Adi.	ArtL	Jad.	Mot.	Mot.E	Pia.	Pia.L	Pip.	Plr.	Plt.	Plt.P	Rec.	She.	Ted.	Vin.
NS-RAFT-Stereo	1.51	**4.14**	24.90	3.62	4.04	9.04	25.81	5.89	14.08	**6.13**	5.54	4.94	39.59	4.96	26.35
Zero-RAFT-Stereo	**1.39**	4.91	**14.27**	**3.26**	**3.68**	**5.69**	**13.73**	**5.22**	**9.53**	7.21	**5.50**	**4.20**	**23.97**	**4.77**	**18.01**

Table 12. Details of Midd-T. We utilize >2px Noc regions in Midd-T (F)

Left | Right, r = 0.1 | Right, r = 0.3 | Right, r = 0.5

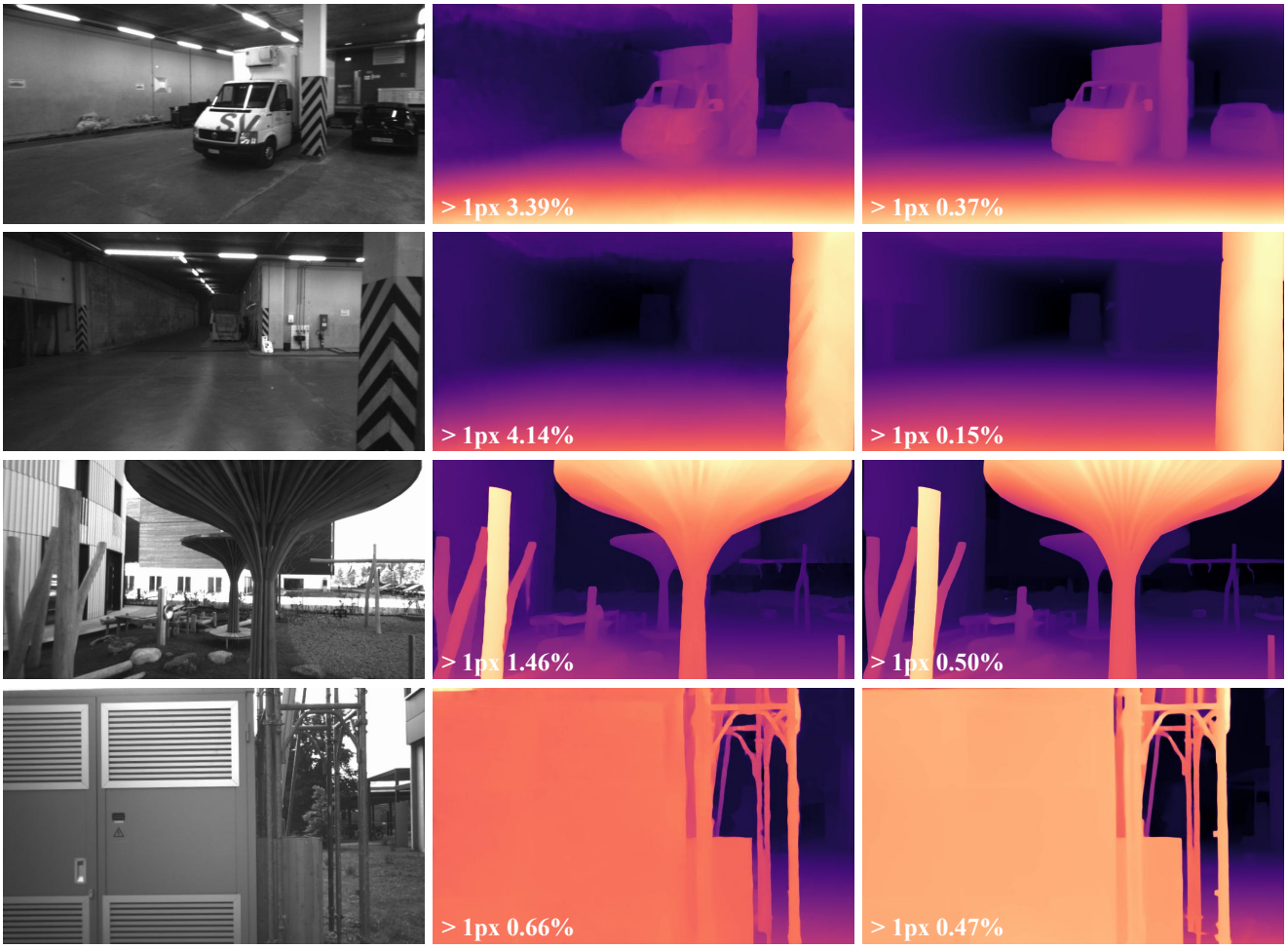
Figure 10. Visualization of AdaMPI [14].

shown in Tab. 11, our Zero-RAFT-Stereo outperforms NS-RAFT-Stereo across all weather conditions except rainy weather, where both models exhibit poor performance, indicating a need for further optimization in such scenarios. Notably, Zero-RAFT-Stereo demonstrates significant improvements under foggy conditions, reducing errors from 18.18% to 8.66% at full resolution and from 3.41% to 1.70% at half resolution. Since foggy scenes typically have low contrast and poor visibility, these results suggest that Zero-RAFT-Stereo is more robust in such challenging conditions. As illustrated in Fig. 13, under extreme weather conditions, NS-RAFT-Stereo struggles to predict large textureless regions, while Zero-RAFT-Stereo successfully reconstructs complete ground surfaces and walls. Moreover, Zero-RAFT-Stereo exhibits superior segmentation of thin, tree-like objects and blurry background regions, highlighting its ability to maintain fine details even in adverse conditions.



Left Image NS-RAFT-Stereo Zero-RAFT-Stereo

Figure 11. Visualization of KITTI.



Left Image NS-RAFT-Stereo Zero-RAFT-Stereo

Figure 12. Visualization of ETH3D.

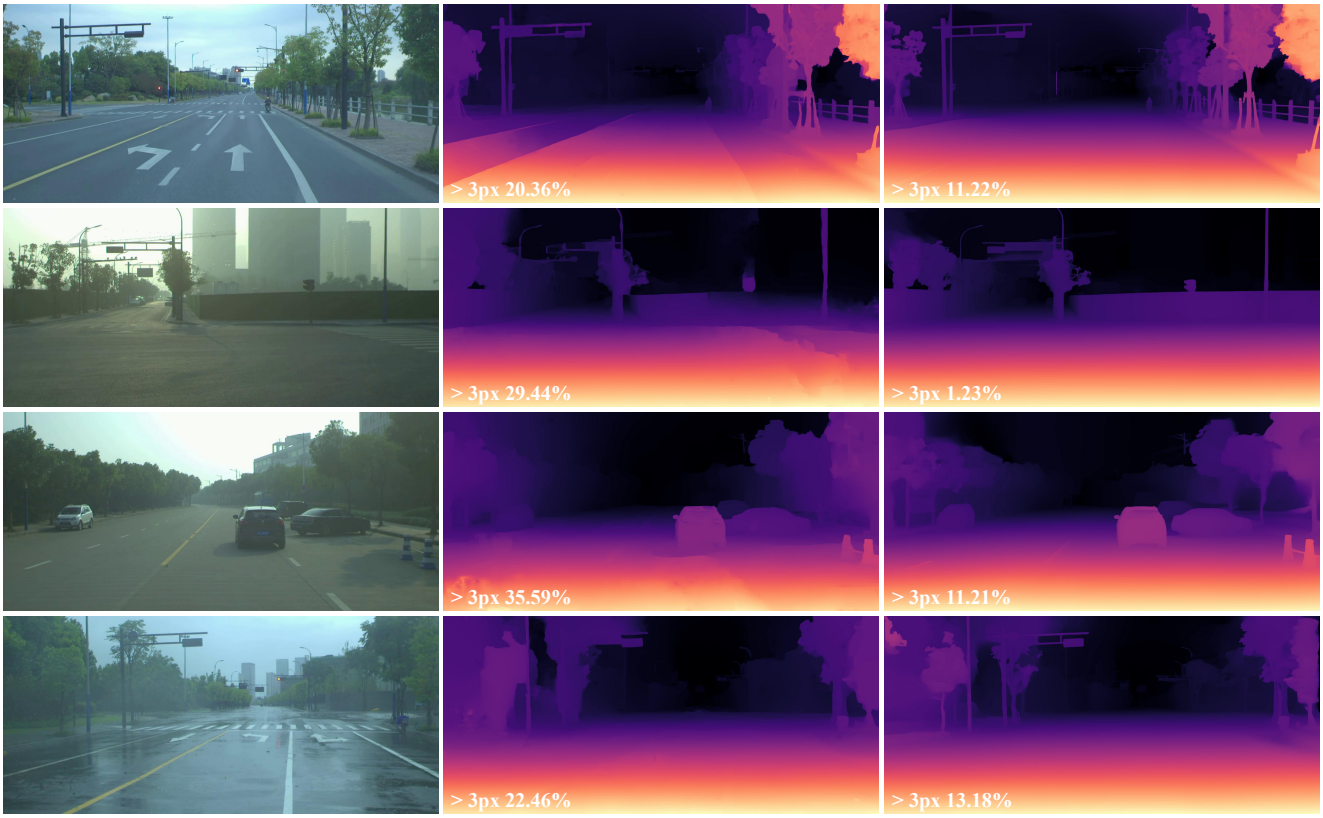


Figure 13. Visualization of DrivingStereo.