This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

RobuSTereo: Robust Zero-Shot Stereo Matching under Adverse Weather

Yuran Wang[1*] Yingping Liang[1*] Yutao Hu[2] Ying Fu[1†]
[1]Beijing Institute of Technology [2]School of Computer Science and Engineering, Southeast University
{wangyuran,liangyingping,fuying}@bit.edu.cn huyutao@seu.edu.cn

Abstract

Learning-based stereo matching models struggle in adverse weather conditions due to the scarcity of corresponding training data and the challenges in extracting discriminative features from degraded images. These limitations significantly hinder zero-shot generalization to out-of-distribution weather conditions. In this paper, we propose ***RobuSTereo****, a novel framework that enhances the zero-shot generalization of stereo matching models under adverse weather by addressing both data scarcity and feature extraction challenges. First, we introduce a diffusion-based simulation pipeline with a stereo consistency module, which generates high-quality stereo data tailored for adverse conditions. By training stereo matching models on our synthetic datasets, we reduce the domain gap between clean and degraded images, significantly improving the models' robustness to unseen weather conditions. The stereo consistency module ensures structural alignment across synthesized image pairs, preserving geometric integrity and enhancing depth estimation accuracy. Second, we design a robust feature encoder that combines a specialized ConvNet with a denoising transformer to extract stable and reliable features from degraded images. The ConvNet captures fine-grained local structures, while the denoising transformer refines global representations, effectively mitigating the impact of noise, low visibility, and weather-induced distortions. This enables more accurate disparity estimation even under challenging visual conditions. Extensive experiments demonstrate that* ***RobuSTereo*** *significantly improves the robustness and generalization of stereo matching models across diverse adverse weather scenarios.*

1. Introduction

Stereo matching is a fundamental task in computer vision that involves estimating the disparity between two input images to derive accurate depth information, which is crucial for various applications, including robotics [57], autonomous driving [21, 35], and augmented reality [51].

While existing methods [12, 24, 49, 50, 60] perform well in most of the normal scenarios, they struggle in adverse weather conditions, such as rain and fog, which are unavoidable in real-world applications. Adverse weather conditions introduce outlier data that significantly degrade zero-shot stereo matching performance and severely impact downstream tasks like autonomous driving. A major factor contributing to this degradation is the substantial domain gap between data captured under normal conditions and in adverse environments. The degradation issue is further exacerbated in zero-shot scenarios, where models are directly applied to unseen weather conditions without fine-tuning. Moreover, images captured in adverse weather often suffer from low visibility, high noise, and strong mirror reflections, making it challenging for stereo matching models to extract reliable features and accurately estimate disparity.

Researchers attempt to address the domain gap problem by expanding datasets to include adverse weather conditions. As shown in Figure 1, some methods use graphics modeling [8] or physics-based simulation [16] to generate stereo datasets. However, prior methods fail to capture complex light effects, such as strong mirror reflections on wet road surfaces, creating a domain gap between synthetic and real images. Others [6, 14] try to collect real-world data in adverse weather, but disparity labels rely on depth sensors like LiDAR, Time-of-Flight (ToF), and Kinect, which perform poorly in bad weather, reducing dataset quality. Additionally, collecting real-world data is extremely time-consuming and labor-intensive. As a result, the scarcity of high-quality, large-scale training datasets under adverse weather conditions remains a significant challenge.

A further limitation arises from the design of stereo matching encoders [9, 23, 42, 49, 59]. Networks pretrained on large image datasets under normal conditions struggle to extract stable and reliable features from low-quality images produced under adverse weather conditions, such as low visibility and high noise. Consequently, the extracted features tend to be significantly noisy, further reducing the accuracy of subsequent matching tasks and making it diffi-

*These authors contributed equally.
†Corresponding author: fuying@bit.edu.cn

Dataset	Image	Pros & Cons
KITTI		Easy to Acquire Real Image Limited Amount No Adverse Senerinos
vKITTI		Easy to Acquire Unrealistic Image Large Amount No Adverse Senerinos
DrivingStereo		Hard to Acquire Real Image Limited Amount Adverse Senerinos
RobuSTereo		Easy to Acquire Realistic Image Unlimited Amount Adverse Senerinos

Figure 1. Comparison of different stereo datasets.

cult for stereo matching models to achieve reliable zero-shot performance in such challenging environments.

In this paper, we propose **RobuSTereo**, a novel stereo matching framework improving zero-shot performance under adverse weather conditions, shown in Figure 2. First, we develop a diffusion-based method for generating stereo training data tailored to challenging weather scenarios. Utilizing a text-to-image diffusion model, our approach generates stereo images under adverse weather conditions using stereo datasets under normal scenarios and diverse text weather prompts. With powerful generative capabilities of diffusion models, we can produce large-scale, high-quality datasets for adverse weather conditions, boosting the generalization of zero-shot stereo matching. Additionally, we incorporate a consistency module to ensure fine-grained consistency between stereo image pairs, which further improves the quality of synthesized training data. On the other hand, to further improve the feature extraction capacity, we propose a stereo matching network that integrates a robust feature encoder. The inclusion of robust feature encoder enhances the feature extraction capability of the network encoder, enabling it to capture more stable features in degraded imaging environments, which, in turn, improves both the accuracy and reliability of zero-shot stereo matching. Experimental results demonstrate that models trained on our synthetic dataset achieve superior zero-shot performance under challenging weather conditions.

To sum up, we make following contributions:

- We propose **RobuSTereo**, a novel stereo matching framework that enhances zero-shot generalization under adverse weather by improving both data generation and feature extraction.
- We introduce a diffusion-based stereo data generation pipeline with a stereo consistency module, which synthesizes high-quality stereo pairs while preserving geometric alignment. This reduces the domain gap and enhances model robustness in adverse weather conditions.
- We design a robust feature encoder that integrates a specialized ConvNet and a denoising transformer, improving feature stability in degraded imaging conditions.

2. Related Work

2.1. Stereo Matching Method

With the successful application of learning-based methods [10, 11, 27–29, 31, 43, 62] across various vision tasks, learning-based stereo matching methods have replaced traditional optimization techniques by leveraging convolutional neural networks (CNNs). The introduction of GC-Net [27] marked a significant shift by employing a 3D convolutional architecture to regularize a 4D cost volume. Building on this success, models such as PSMNet [9], Gwc-Net [22], and GANet [55] have further improved accuracy. To enhance efficiency, cascade-based methods like CFNet [39] were introduced. IGEV [49] proposed a novel module to capture non-local geometric and contextual information. Additionally, StereoBase [23] provides a strong baseline model, further improving stereo network performance. Recently, StereoAnything [24], pre-trained on both simulated and real datasets, achieves impressive zero-shot results. However, these methods are designed for normal environments and struggle to handle image degradation under harsh conditions. Therefore, improving model robustness for such challenging scenarios is essential.

2.2. Datasets and Data Generation Method

Stereo datasets can be broadly divided into real-world and synthetic datasets. Among real-world datasets, KITTI12 [18] and KITTI15 [33] provide 400 training pairs of outdoor stereo images with sparse disparity labels from LiDAR measurements. Recently, large-scale datasets such as HRWSI [48] and IRS [46] have emerged. While these datasets improve data volume, most are captured under normal lighting, lacking training data for complex scenes. Some researchers [6, 14, 17, 52] have collected data in adverse weather, but the quality is limited by the inaccuracy of LiDAR and other equipment. Meanwile, synthetic datasets like SceneFlow [32] and vKITTI [8] use computer graphics (CG) techniques to generate stereo images with precise disparity maps. vKITTI [8] attempts to improve the diversity of the dataset by simulating different weather conditions such as rainy. However, CG-based methods [8, 32, 61] struggle to replicate real-world complexity, leading to domain gaps that limit their real-world applicability.

Image Diffusion. Latent Diffusion Models (LDMs) [36] reduce computational costs by performing denoising in latent space. Conditioning approaches include cross-attention [2, 7, 15], segmentation masks as tokens [3], and other strategies [5, 30, 58] enabling visual generation from diverse inputs such as text, images, depth, and videos. Beyond image

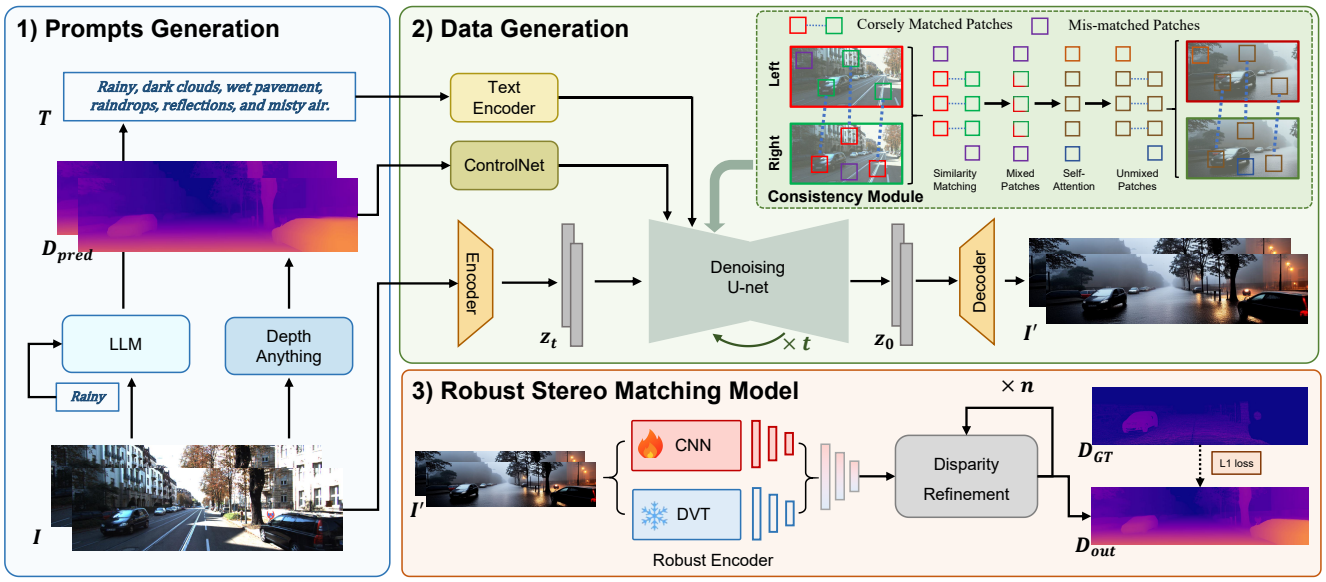


Figure 2. RobuSTereo Framework for Stereo Matching under Adverse Weather Conditions. The framework consists of three components: (1) Prompts Generation, where depth and weather prompts are generated using a Large Language Model (LLM) and depth estimation to describe adverse weather conditions; (2) Data Generation, where the prompts and depth map guide a ControlNet and denoising U-Net to produce a weather-affected synthetic image; and (3) Robust Stereo Matching Model, where the synthetic image is used to train stereo matching networks with L1 loss for improved stereo matching under challenging weather.

synthesis, diffusion models excel at optical flow [38] and monocular depth estimation [26, 44, 56]. Our method leverages LDMs to generate diverse stereo pairs under complex weather, enriching stereo datasets for adverse conditions.

3. Method

In this section, we first present the motivation of our proposed method. Next, we introduce data generation pipeline and the robust stereo matching network in detail. An illustration of the proposed framework is provided in Figure 2.

3.1. Motivation and Formulation

Accurate zero-shot stereo matching in adverse weather, such as rain, fog, and snow, is not only essential but also highly challenging, as such conditions are unavoidable in real life and greatly reduce image quality, which directly affects stereo matching performance. The lack of large-scale, high-quality datasets for adverse weather further worsens these challenges, making it difficult to achieve reliable stereo matching. Existing datasets, such as DrivingStereo [52], are limited in diversity and scale, and sensors like LiDAR struggle to provide accurate disparity under such conditions. Furthermore, most stereo matching networks lack robust encoders capable of handling the degraded image quality caused by adverse weather.

To address these challenges, we propose **RobuSTereo**, which generates stereo data under various adverse weather conditions and introduces a more powerful stereo matching encoder. Our method enables scalable data generation with accurate disparity maps, offering greater diversity and ease of acquisition than existing datasets. Furthermore, we optimize stereo matching networks by introducing a robust feature encoder, enhancing feature extraction stability and improving performance in complex adverse environments.

3.2. Stereo Data Generation for Adverse Weather

A training pair $(\mathbf{I}_R, \mathbf{I}_L, \mathbf{D})$ of a stereo dataset consists of two parts: stereo images $\mathbf{I}_R, \mathbf{I}_L \in \mathbb{R}^{3\times H\times W}$ and the disparity map $\mathbf{D} \in \mathbb{R}^{H\times W}$ (typically the disparity of the left view). Most of the existing image datasets are data under normal weather domain $(\mathbf{I}_R, \mathbf{I}_L, \mathbf{D})_{norm}$. We hope to transfer them to adverse environments domain $(\mathbf{I}'_R, \mathbf{I}'_L, \mathbf{D})_{adv}$ through style transfer, where $\mathbf{I}'$ represents image under adverse weather. Then we use these data pairs to train a more robust stereo matching network. Therefore, the task can be formulated as $(\mathbf{I}'_R, \mathbf{I}'_L, \mathbf{D}) = \mathcal{T}(\mathbf{I}_R, \mathbf{I}_L, \mathbf{D})$, where $\mathcal{T}$ stands for the desired transformation method.

Diffusion-based Stereo Image Generation. IDMs has obvious advantages in image style transfer and image editing. Therefore, we introduce the diffusion method using ControlNet to achieve stereo image generation.

First, we generate descriptors for the target domain. With the powerful understanding and generalization capabilities of the large language models (LLMs), we use them to help us generate powerful prompt keywords. Taking the conversion of an image pair to a rainy day as an example,

we input the source image into LLMs [1] and ask it to generate prompt words under the corresponding conditions according to the image content, like *describe a rainy day of this image in a few keywords*. We collect the prompts T = *Rainy, dark clouds, wet pavement, raindrops, reflections, and misty air* for subsequent image style transfer.

Then, we use Depth2Image ControlNet [58] as a guidance for image content to ensure that the generated image is consistent with the original image in depth information. In this way, the generated image can match the original disparity GT ($\mathbf{D}_{GT}$). Considering that $\mathbf{D}_{GT}$ of some stereo datasets is sparse and cannot be used as the input of the controlnet, we first use the depth estimation network [54] to predict the depth map $\mathbf{D}_{pred}$ of the image and use it as the control input of the controlnet. Since the existing images are generally image pairs in a clear and simple environment, the image depth obtained in this step has a high credibility. The conditioning process is:

$$\mathbf{D}_{pred} = (\mathcal{D}(\mathbf{I}_R), \mathcal{D}(\mathbf{I}_L)), \tag{1}$$

$$\mathbf{c} = \mathcal{F}_{CtrlNet}(\mathbf{z}_t, \mathbf{D}_{pred}, T), \tag{2}$$

where $\mathcal{F}_{CtrlNet}(\cdot)$ represents Depth2Image ControlNet [58], $\mathcal{D}(\cdot)$ is DepthAnythingV2 [54], T is relevant text prompts, t denotes the t-th step of the backward diffusion process and c is conditioning feature, which is then utilized in the Stable Diffusion (SD) generation process

$$\mathbf{I}' = \mathcal{F}_{SD}(\mathbf{z}_t, T|\mathbf{c}), \tag{3}$$

$\mathbf{I}'$ is for the generated image for target domain, and $\mathcal{F}_{SD}(\cdot)$ stands for the generation process of pretrained Stable Diffusion model conditioned by a ControlNet. Diffusion-base methodology enables the generation of a wide range of stereo image pairs that maintain the underlying scene structure while exhibiting varied visual characteristics, simulating different adverse weather conditions. Example sythetic images are shown in Figure 3. Theoretically, we can generate an unlimited amount of relevant data.

Coherence-Enhanced Consistency Module. Diffusion-based methods have clear advantages in generating detailed and diverse content. However, when generating stereo images, excessive diversity in the generated content can result in the left and right images not corresponding properly. Such inconsistencies make generated images unsuitable for training stereo networks. To address this issue, we introduce specific consistency constraints during the generation process to enhance the content similarity between the left and right images. This ensures that the generated stereo images are more aligned and suitable for stereo network training.

Inspired by video editing methods, we propose a Disparity Fusion Method (DFM) to improve the consistency of generated images. Given the input image patches $\mathbf{P}$, DFM first divides the patches into a source set (*src*) and a destination set (*dst*), and computes the patch-wise similarity

Figure 3. Visualization of generated data under adverse weather using our proposed RobuSTereo from stereo images under normal conditions. From top to bottom: RGB stereo images, Snowy, Foggy and Rainy.

between these two sets. The similarity between two patches is calculated based on their disparity similarity and image similarity. Each patch in the *src* set is linked to its most similar patch in the *dst* set, and the top-*n* most similar patch pairs are fused. The formulation is expressed as

$$\mathbf{E} = Match(src, dst, n), \tag{4}$$

$$\mathbf{P}_m = \mathcal{M}(\mathbf{P}, \mathbf{E}), \quad \mathbf{P}_u = \mathcal{U}(\mathbf{P}_m, \mathbf{E}), \tag{5}$$

$Match(\cdot)$ represents the process of finding the top-*n* most similar patch pairs, while $\mathcal{M}(\cdot)$ and $\mathcal{U}(\cdot)$ denote the patch fusion and un-fusion processes, respectively, based on the matching map $\mathbf{E}$. In the context of diffusion-based stereo image generation, the input and output patches are denoted as $\mathbf{P}_{in}, \mathbf{P}_{out} \in \mathbb{R}^{2\times N\times C}$, where N is the number of patches, and C is the feature dimension. Based on Equation (4), we perform the fusion process as

$$\mathbf{P}_m = \mathcal{M}(\mathbf{P}_{in}, Match(\mathbf{P}_{in}^{left}, \mathbf{P}_{in}^{right}, n)), \tag{6}$$

$$\mathbf{P}_{out} = \mathcal{U}(Attn(\mathbf{P}_m), Match(\mathbf{P}_{in}^{left}, \mathbf{P}_{in}^{right}, n)). \tag{7}$$

The fused patches $\mathbf{P}_m$ are then passed through a self-attention mechanism, denoted as $Attn(\cdot)$, to identify and refine consistent features. This approach significantly improves the multi-view consistency of the generated stereo images, ensuring that the left and right views are well-aligned and suitable for stereo network training.

3.3. Stereo Networks with Robust Feature Encoder

In adverse weather, RGB image imaging is often affected by noise and low visibility, which poses significant challenges for stereo matching. Traditional stereo matching methods typically rely on simple encoders pre-trained on ImageNet [13] or other large-scale RGB datasets as their feature extraction modules. However, these encoders are not

Networks	Publication	Rainy		Sunny		Foggy		Cloudy		Overall	
		EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓
PSMNet [9]	CVPR'18	20.864	50.862	3.668	27.499	19.562	58.040	4.437	30.989	12.133	41.848
CFNet [39]	CVPR'21	4.211	23.558	2.178	15.060	3.442	25.910	3.392	23.277	3.306	21.951
GwcNet [22]	CVPR'19	6.211	48.851	2.957	23.896	4.721	43.887	3.760	29.950	4.412	36.646
COEX [4]	IROS'21	4.531	28.959	2.490	17.264	3.093	22.694	2.995	21.849	3.277	22.692
FADNet++ [45]	ICRA'20	2.852	24.234	1.999	15.019	2.214	19.566	2.063	15.985	2.282	18.701
CasStereo [19]	CVPR'20	5.013	33.692	3.612	22.732	4.143	31.444	3.863	26.123	4.158	28.498
IGEV [49]	CVPR'23	1.879	10.955	1.215	5.080	1.253	6.582	1.077	4.196	1.356	6.704
StereoBase [23]	ArXiv'23	1.695	8.610	1.198	5.023	1.224	5.980	1.090	4.281	1.302	5.974
NMRF [20]	CVPR'24	2.490	12.073	1.046	3.674	1.089	4.337	0.976	3.162	1.401	5.811
Selective-IGEV [47]	CVPR'24	1.178	5.397	1.103	4.296	2.172	13.660	1.127	4.820	1.395	7.043
StereoAnything [24]	ArXiv'24	1.144	5.395	0.984	2.972	1.134	4.821	0.906	2.273	1.042	3.865
LightStereo [25]	ICRA'25	1.105	4.846	1.078	3.609	1.155	4.927	1.014	3.046	1.088	4.107
MonSter [12]	CVPR'25	1.153	5.335	1.030	3.509	1.152	5.275	0.987	3.180	1.081	4.325
RobuSTereo (Ours)	-	**0.973**	**1.939**	**0.785**	**1.493**	**0.853**	**1.610**	**0.735**	**1.349**	**0.836**	**1.598**

Table 1. Comparison of model performance on DrivingStereo [52] with stereo matching models. "Ours" refers to our model trained on our generated dataset. All models are validated on the DrivingStereo weather subset. Most methods are trained on SceneFlow [32], except StereoAnything [24], LightStereo [25], and MonSter [12], which are trained on mixed datasets combining synthetic and real-world data.

specifically designed to handle the degraded image quality encountered in adverse weather conditions, making them difficult to process such challenging scenes.

In this paper, in order to improve the stability of feature extraction by the encoder, we propose a robust feature encoder by combining Denosing-Vision Transformer (DVT) [53] and VGG19 [41]. DVT captures richer and more robust high-dimensional feature representations, including both semantic and high-level information, while VGG focuses on extracting detailed pixel-wise features. Compared to traditional CNN-based encoder or ViT encoder, DVT can reduce the noise of features at the feature level to obtain more stable features, solving the problem of poor image quality in complex weather conditions.

Formally, given a pair of left and right images, $\mathbf{I}_L, \mathbf{I}_R \in \mathbb{R}^{H\times W\times 3}$, we first utilize the well-established VGG19 [41] to extract feature. VGG19 [41] generates multi-scale pyramid features, producing detailed feature maps at different resolutions: $f_c^{(i)} \in \mathbb{R}^{C_i\times \frac{H}{i}\times \frac{W}{i}}$, where $i \in \{4, 8, 16\}$ represents the scaling factor. These multi-scale features provide a rich hierarchical representation of the input images. To further enhance the feature extraction process, we incorporate DVT, which generates an additional high-dimensional and robust feature map, $f_c^{(32)} \in \mathbb{R}^{C_{32}\times \frac{H}{32}\times \frac{W}{32}}$, capturing high-level semantic and contextual information, which plays a critical role in improving the robustness and accuracy of stereo matching, especially in complex scenarios.

Finally, the disparity refinement network is designed following on the method proposed in [23], which enhances depth estimation by refining the initial disparity maps iteratively. The integration of advanced feature extraction improves the prediction accuracy under adverse weather.

4. Experiments

In this section, we introduce the datasets, implementation details, and conduct zero-shot comparisons across different datasets. Ablation studies validate the effectiveness of our main components, with additional analysis on visual point-clouds provided in the supplementary materials.

4.1. Experimental Setup

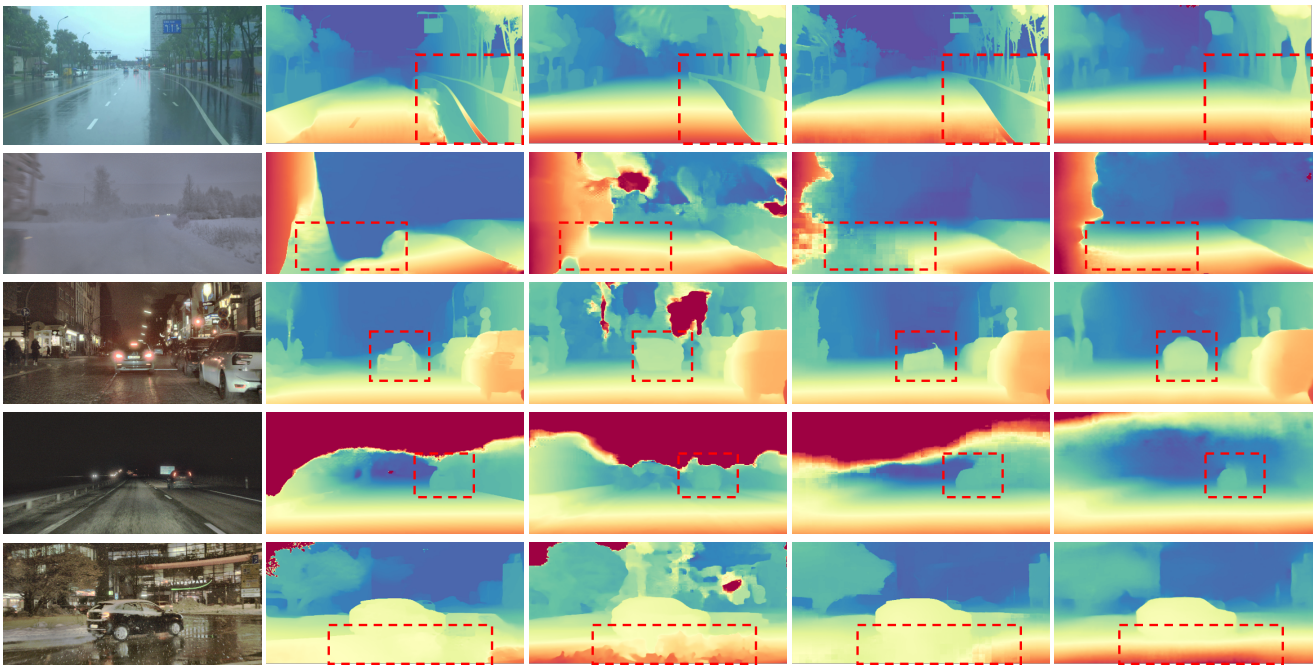
Datasets. In our experiments, we utilize several datasets for training and evaluation. **SceneFlow** [32], **KITTI 2012** [18], **KITTI 2015** [33], and **Virtual KITTI V2** (**vKITTI**) [8] is used for training. **DrivingStereo** [52] and **SeeingThrough-Fog** [6] serve as test sets for performance evaluation under different adverse weather conditions for both quantitative evaluations and visualization. We provide more detailed dataset information in the supplementary material.

Image Generation Configurations. We use the Diffusers library to modify and deploy Stable Diffusion [36] (SD) and ControlNet [58]. SD (version 1.5) acts as the image generator, with a DDIM scheduler and 50 sampling steps. For depth-conditioned generation, we use ControlNet [58] (version *control_v11f1p_sd15_depth*) and Depth-Anything V2 [54] for monocular depth estimation. Source images from KITTI [18, 33] and vKITTI [8] are used to generate stereo data for adverse weather. The synthetic dataset is named Robust-Dataset (**RST-Dataset**).

Stereo Matching Models. We evaluate our framework against traditional and advanced methods, including PSMNet [9], CFNet [39], GwcNet [22], COEX [4], FADNet++ [45], CasStereo [19], IGEV [49], StereoBase [23], NMRF [20], Selective-IGEV [47], LightStereo [25],

Networks	Datasets	Snow		Rain		Dense Fog		Light Fog		Overall	
		EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓
StereoAnything [24]	MIX	4.265	28.734	3.440	21.955	6.055	27.290	2.973	16.170	4.204	26.526
LightStereo [25]	MIX	3.894	30.104	3.034	24.836	5.904	31.083	2.622	19.118	3.853	28.431
StereoBase [23]	SceneFlow [32]	6.567	33.623	5.150	28.592	6.347	30.958	3.071	20.955	6.007	31.434
	KITTI [33]	8.001	29.802	5.918	21.333	5.533	22.101	2.801	13.634	6.976	26.449
	vKITTI [8]	5.519	34.634	4.447	28.408	6.621	33.801	3.264	22.450	5.259	32.573
	RST-Dataset	7.059	27.532	4.516	18.107	5.601	20.429	2.438	10.301	6.158	24.026
RobuSTereo (Ours)	RST-Dataset	**3.409**	**23.438**	**2.577**	**15.656**	**5.317**	**20.128**	**2.120**	**10.074**	**3.359**	**20.881**

Table 2. Comparison of zero-shot performance on SeeingThroughFog [6]. StereoAnything [24], LightStereo [25] are trained on mixed synthetic and real-world datasets. Stereobase [23] model are trained on four different datasets.



Left Image | StereoBase-SF [23] | StereoBase-KITTI [23] | StereoAnything [24] | Ours

Figure 4. Qualitative results under adverse conditions. Our method produces more accurate and consistent disparity maps, preserving fine details and reducing artifacts compared to other approaches, as highlighted in the red boxes.

StereoAnything [24] and MonSter [12]. For dataset-level experiments, we focus on PSMNet [9], IGEV [49], and StereoBase [23], which are trained on various datasets and evaluated on the DrivingStereo [52] dataset.

Training Details. All models are trained on the SceneFlow dataset and our **RST-dataset**. The batch size is uniformly set to 8, and each model undergoes the same number of training steps across the various datasets. Models are then evaluated on adverse weather conditions from the DrivingStereo [52] and SeeingThroughFog [6] dataset to assess their generalization performance.

Evaluation Metrics. We evaluate performance using End-Point Error (EPE), which measures average pixel-level disparity error, and the D1 metric, which quantifies the percentage of disparity outliers exceeding 3 pixels or 5% of the true value, critical for assessing models' performance.

4.2. Quantitative Results

Comparison with Stereo Matching Methods. Comparison results are presented in Table 1 and Table 2, where we compare state-of-the-art stereo matching methods on adverse weather datasets. On the DrivingStereo [52] dataset, our method achieved the best results across all weather conditions, particularly in challenging scenarios such as rainy days, demonstrating its effectiveness in addressing stereo matching issues under adverse weather.

Networks	Datasets	Rainy		Sunny		Foggy		Cloudy		Overall	
		EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓	EPE ↓	D1 ↓
PSMNet [9]	SceneFlow [32]	20.864	50.862	3.668	27.499	19.562	58.040	4.437	30.989	12.133	41.848
	KITTI [33]	1.632	6.077	1.241	3.687	1.208	5.107	0.973	2.648	1.264	4.380
	vKITTI [8]	1.579	9.007	1.418	6.404	1.342	8.439	1.036	3.463	1.344	6.829
	RST-Dataset (Ours)	**1.208**	**4.723**	**1.058**	**3.317**	**1.076**	**4.722**	**0.912**	**2.367**	**1.064**	**3.783**
IGEV [49]	SceneFlow [32]	1.879	10.955	1.215	5.080	1.253	6.582	1.077	4.196	1.356	6.704
	KITTI [33]	1.456	4.915	**0.838**	**2.026**	0.917	**1.988**	**0.822**	**2.046**	1.009	2.792
	vKITTI [8]	1.533	7.529	1.151	3.679	1.208	5.471	0.992	3.038	1.221	4.930
	RST-Dataset (Ours)	**1.082**	**3.467**	0.920	2.308	**0.902**	2.184	0.867	2.140	**0.943**	**2.525**
StereoBase [23]	SceneFlow [32]	1.695	8.610	1.198	5.023	1.224	5.980	1.090	4.281	1.302	5.974
	KITTI [33]	1.150	2.857	0.840	1.947	0.920	2.408	0.820	1.904	0.927	2.279
	vKITTI [8]	1.244	6.949	1.108	4.005	1.288	6.399	0.992	3.273	1.159	5.157
	RST-Dataset (Ours)	**1.003**	**2.221**	**0.834**	**1.834**	**0.867**	**2.383**	**0.797**	**1.764**	**0.875**	**2.050**

Table 3. Comparison of zero-shot performance on DrivingStereo [52] dataset using different datasets. We train three representative models, including PSMNet [9], IGEV [49], and StereoBase [23]. Among them, StereoBase represents the state-of-the-art model.

Additionally, in Table 2, we evaluate on the extreme adverse weather dataset SeeingThroughFog [6], comparing powerfull zero-shot models StereoAnything [24] and LightStereo [25]. To validate the effectiveness of our dataset, we fine-tuned StereoBase [23] using various datasets. The results demonstrate that our method largely improves stereo matching performance under adverse weather. Notably, even training existing models using only the RST-dataset, without additional model enhancements, yields substantial performance gains in adverse weather conditions.

Comparison with Other Stereo Datasets. As shown in Table 3, we validated performance on generated data, showing that the model trained on RST-dataset achieves results comparable to state-of-the-art datasets like SceneFlow [32], KITTI [33], and vKITTI [8]. While performance across datasets is similar in normal scenarios, the RST-dataset demonstrates superior adaptability in adverse weather conditions, such as rain and fog, highlighting its effectiveness in enhancing generalization for real-world applications.

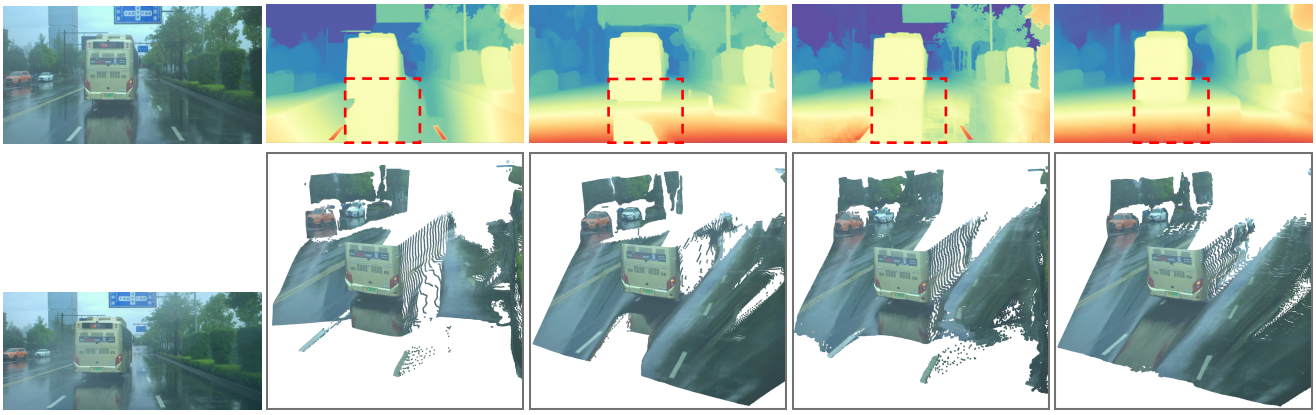
4.3. Qualitative Results

Disparity Maps in Adverse Weather. As shown in Figure 4, RobuSTereo outperforms other models, particularly under challenging conditions. In rainy weather, the mirror effect from wet surfaces creates challenges for stereo matching algorithms, often leading to errors in estimating ground disparity. This issue is critical for applications like autonomous driving, where inaccurate depth perception can impede obstacle detection and compromise safety. In contrast, models trained on our dataset excel on specular surfaces, providing more accurate road depth information and ensuring improved performance in such conditions.

Additionally, Figure 6 compares the point cloud reconstructions from the Driving Stereo (Rainy) dataset. The improved reconstructions demonstrate the robustness of our approach in challenging weather conditions, highlighting the practical benefits of using our dataset to develop more reliable stereo vision systems for real-world applications.

Figure 5. Comparison of Feature Maps from Different Encoders. Feature maps are visualized using PCA [40]. Compared to other methods, our feature maps exhibit better consistency and reduced noise, improving the accuracy of subsequent stereo matching.

Feature Maps in Adverse Weather. Figure 5 illustrates the feature representations extracted by different encoders under adverse weather conditions. StereoBase [23] use MobileNetV2 [37] as encoder. DinoV2 [34] denotes a standard ViT encoder. Due to the degraded image quality, noticeable noise is observed in the extracted features in MobileNetV2 [37] and DinoV2 [34], which impacts the subsequent matching process. Notably, our encoder effectively reduces the noise in the feature representations, enhancing the robustness of feature extraction. This improvement leads to better performance and stability of the stereo matching model in challenging weather scenarios.



Left/Right Image | StereoBase-SF [23] | StereoBase-KITTI [23] | StereoAnything [24] | Ours

Figure 6. Qualitative visualization of point cloud results under adverse conditions. Point clouds generated by other methods show significant artifacts on wet surfaces due to specular reflections. In contrast, our method effectively mitigates these issues, producing accurate point clouds with improved geometric consistency. A set of full-view point cloud videos are provided in the supplementary material.

Experiments	Methods	Driving Stereo	
		EPE ↓	D1 ↓
Consistency Module	Off	1.308	5.997
	On	**0.875**	**2.050**
Source Data	vKITTI [8]	1.039	4.689
	KITTI [33]	**0.875**	**2.050**
Network Encoder	MobileNetV2 [37]	0.875	2.050
	Dinov2 [34]	0.852	1.793
	Robust Encoder	**0.836**	**1.598**

Table 4. Ablation experiments. Without changing other conditions such as prompt, we removed the consistency constraint module/changed the data source and generate the same number of images, using StereoBase [23] as base model, and finally tested the comprehensive performance on DrivingStereo [52].

4.4. Ablation Study

Coherence-Enhanced Consistency Module. In the ablation study, we compare the proposed consistency module under same data generation config. As shown in Table 4, generating images without the consistency module leads to a degradation in subsequent training results. Without proper correspondence between the generated images, networks struggle to match the images correctly, which introduces confusion during the matching process.

Data Source. We evaluate the performance of different data sources using KITTI and vKITTI. While vKITTI offers the advantage of accurate and dense disparity ground truth (GT) and a larger dataset compared to KITTI, our experiments reveal that the performance of generated stereo images using vKITTI does not outperform that of KITTI. This discrepancy may be attributed to the relatively simple image textures in vKITTI, which differ from real-world scenes, creating a domain gap in the generated images.

Robust Feature Encoder. We evaluated the performance of various encoders integrated into the stereo matching model under adverse weather conditions. All experiments are conducted using the same training pipeline. The results demonstrate that our robust encoder delivers significant improvements in accuracy, which stems from encoder's ability to extract consistent and robust features, even when image quality is degraded due to adverse weather.

5. Conclusion

In this work, we propose RobuSTereo, a new framework to improve zero-shot stereo matching performance in adverse weather by addressing challenges in both data and models. First, we introduce a diffusion-based data generation pipeline to create high-quality stereo datasets under adverse weather, ensuring the generated stereo image pairs are realistic and consistent. We also design a coherence-enhanced consistency module to further improve the consistency of the generated images. Second, we strengthen the feature extraction of stereo matching models by using a robust encoder, which helps the model handle poor image quality caused by bad weather. Both quantitative and qualitative results show that RobuSTereo greatly improves the accuracy and reliability of stereo matching models, reducing errors caused by issues like reflections and low visibility. Our results show the potential of using advanced data generation methods to improve stereo matching, especially for scenarios that are not well-represented in existing datasets.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFC3300704), the National Natural Science Foundation of China (62331006, 62171038, and 62088101), and the Fundamental Research Funds for the Central Universities.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint*, 2023. 4

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2

[3] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 2

[4] Antyanta Bangunharcana, Jae Won Cho, Seokju Lee, In So Kweon, Kyung-Soo Kim, and Soohyun Kim. Correlate-and-excite: Real-time stereo matching via guided cost volume excitation. In *Proceedings of the IEEE Int. Conference on Intelligent Robots and Systems*, 2021. 5

[5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 2

[6] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020. 1, 2, 5, 6, 7

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[8] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020. 1, 2, 5, 6, 7, 8

[9] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1, 2, 5, 6, 7

[10] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 2

[11] Linwei Chen, Ying Fu, Lin Gu, Dezhi Zheng, and Jifeng Dai. Spatial frequency modulation for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):1–18, 2025. 2

[12] Junda Cheng, Longliang Liu, Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Yong Deng, Jinliang Zang, Yurui Chen, Zhipeng Cai, and Xin Yang. Monster: Marry monodepth to stereo unleashes power. *arXiv preprint*, 2025. 1, 5, 6

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2009. 4

[14] Zaid A El-Shair, Abdalmalek Abu-raddaha, Aaron Cofield, Hisham Alawneh, Mohamed Aladem, Yazan Hamzeh, and Samir A Rawashdeh. Sid: Stereo image dataset for autonomous driving in adverse conditions. In *NAECON 2024-IEEE National Aerospace and Electronics Conference*, pages 403–408. IEEE, 2024. 1, 2

[15] Oran Gafni, Adam Polyak, and Yaniv Nechemia Taigman. Scene-based text-to-image generation with human priors, 2024. US Patent App. 18/149,542. 2

[16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4340–4349, 2016. 1

[17] Yunyi Gao, Qiankun Liu, Lin Gu, and Ying Fu. Grayscale-assisted rgb image conversion from near-infrared images. *Tsinghua Science and Technology*, 30(5):2215–2226, 2025. 2

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2, 5

[19] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2020. 5

[20] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2024. 5

[21] Wenxuan Guo, Yingping Liang, Zhiyu Pan, Ziheng Xi, Jianjiang Feng, and Jie Zhou. Camera-lidar cross-modality gait recognition. *arXiv preprint*, 2024. 1

[22] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. 2, 5

[23] Xianda Guo, Juntao Lu, Chenming Zhang, Yiqi Wang, Yiqun Duan, Tian Yang, Zheng Zhu, and Long Chen. Openstereo: A comprehensive benchmark for stereo matching and strong baseline. *arXiv preprint*, 2023. 1, 2, 5, 6, 7, 8

[24] Xianda Guo, Chenming Zhang, Youmin Zhang, Dujun Nie, Ruilin Wang, Wenzhao Zheng, Matteo Poggi, and Long Chen. Stereo anything: Unifying stereo matching with large-scale mixed data. *arXiv preprint*, 2024. 1, 2, 5, 6, 7, 8

[25] Xianda Guo, Chenming Zhang, Youmin Zhang, Wenzhao Zheng, Dujun Nie, Matteo Poggi, and Long Chen. Lightstereo: Channel boost is all your need for efficient 2d cost aggregation. *arXiv preprint*, 2024. 5, 6, 7

[26] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 3

[27] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 66–75, 2017. 2

[28] Hesong Li and Ying Fu. Fcdfusion: A fast, low color deviation method for fusing visible and infrared image pairs. *CVM*, 11(1):195–211, 2025.

[29] Miaoyu Li, Ying Fu, Tao Zhang, Ji Liu, Dejing Dou, Chenggang Yan, and Yulun Zhang. Latent diffusion enhanced rectangle transformer for hyperspectral image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[30] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2024. 2

[31] Qiankun Liu, Yuqi Jiang, Zhentao Tan, Dongdong Chen, Ying Fu, Qi Chu, Gang Hua, and Nenghai Yu. Transformer based pluralistic image completion with reduced information loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[32] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 2, 5, 6, 7

[33] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 2, 5, 6, 7, 8

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint*, 2023. 7, 8

[35] Mur-Artal R Tardós JD Orb. Slam2: an open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 1

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 5

[37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 7, 8

[38] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Proceedings of Advances in Neural Information Processing Systems*, 36, 2024. 3

[39] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2021. 2, 5

[40] Jonathon Shlens. A tutorial on principal component analysis. *arXiv preprint*, 2014. 7

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of International Conferenceon Learning Representations*, 2015. 5

[42] Jian Su, Shiang Mao, and Wei Zhuang. Aoyolo algorithm oriented vehicle and pedestrian detection in foggy weather. *Chinese Journal of Electronics*, 34(2):661–672, 2025. 1

[43] Ye Tian, Ying Fu, and Jun Zhang. Transformer-based undersampled single-pixel imaging. *Chinese Journal of Electronics*, 32(5):1151–1159, 2023. 2

[44] Fabio Tosi, Pierluigi Zama Ramirez, and Matteo Poggi. Diffusion models for monocular depth estimation: Overcoming challenging conditions. *arXiv preprint*, 2024. 3

[45] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. Fadnet: A fast and accurate network for disparity estimation. In *Proceedings of the IEEE Int. Conference on Robotics and Automation*, pages 101–107. IEEE, 2020. 5

[46] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. pages 1–6, 2021. 2

[47] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 5

[48] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 611–620, 2020. 2

[49] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 1, 2, 5, 6, 7

[50] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *Proceedings of European Conference on Computer Vision*, pages 214–230. Springer, 2022. 1

[51] Aimin Yang, Chunying Zhang, Yongjie Chen, Yunxi Zhuansun, and Huixiang Liu. Security and privacy of smart home systems based on the internet of things and stereo match-

ing algorithms. *IEEE Internet of Things Journal*, 7(4):2521–2530, 2019. 1

[52] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 899–908, 2019. 2, 3, 5, 6, 7, 8

[53] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas J. Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Dvt: Denoising vision transformers. 2024. 5

[54] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint*, 2024. 4, 5

[55] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. 2

[56] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 11852–11861, 2024. 3

[57] Guoxuan Zhang, Jin Han Lee, Jongwoo Lim, and Il Hong Suh. Building a 3-d line-based map using stereo slam. *IEEE Transactions on Robotics*, 31(6):1364–1377, 2015. 1

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 4, 5

[59] Tao Zhang, Ying Fu, and Jun Zhang. Deep guided attention network for joint denoising and demosaicing in real image. *Chinese Journal of Electronics*, 33(1):303–312, 2024. 1

[60] Jin Zheng, Botao Jiang, Wei Peng, and Qiaohui Zhang. Multi-scale binocular stereo matching based on semantic association. *Chinese Journal of Electronics*, 33(4):1010–1022, 2024. 1

[61] Tianyu Zhu, Hesong Li, and Ying Fu. Trim-sod: A multi-modal, multi-task, and multi-scale spacecraft optical dataset. *Space: Science & Technology*, 2025. 2

[62] Yunhao Zou, Ying Fu, Tsuyoshi Takatani, and Yinqiang Zheng. Eventhdr: From event to high-speed hdr videos and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2