Boosting Zero-shot Stereo Matching Using Large-Scale Mixed Images Sources in the Real World

Yuran Wang, **Yingping Liang**, **Ying Fu**
Beijing Institute of Technology
{wangyuran, liangyingping, fuying}@bit.edu.cn

Abstract

Stereo matching methods rely on dense pixel-wise ground truth labels, which are laborious to obtain, especially for real-world datasets. The scarcity of labeled data and domain gaps between synthetic and real-world images also pose notable challenges. In this paper, we propose a novel framework, BooSTer, that leverages both vision foundation models and large-scale mixed image sources, including synthetic, real, and single-view images. First, to fully unleash the potential of large-scale single-view images, we design a data generation strategy combining monocular depth estimation and diffusion models to generate dense stereo matching data from single-view images. Second, to tackle sparse labels in real-world datasets, we transfer knowledge from monocular depth estimation models, using pseudo-mono depth labels and a dynamic scale- and shift-invariant loss for additional supervision. Furthermore, we incorporate vision foundation model as an encoder to extract robust and transferable features, boosting accuracy and generalization. Extensive experiments on benchmark datasets demonstrate the effectiveness of our approach, achieving significant improvements in accuracy over existing methods, particularly in scenarios with limited labeled data and domain shifts.

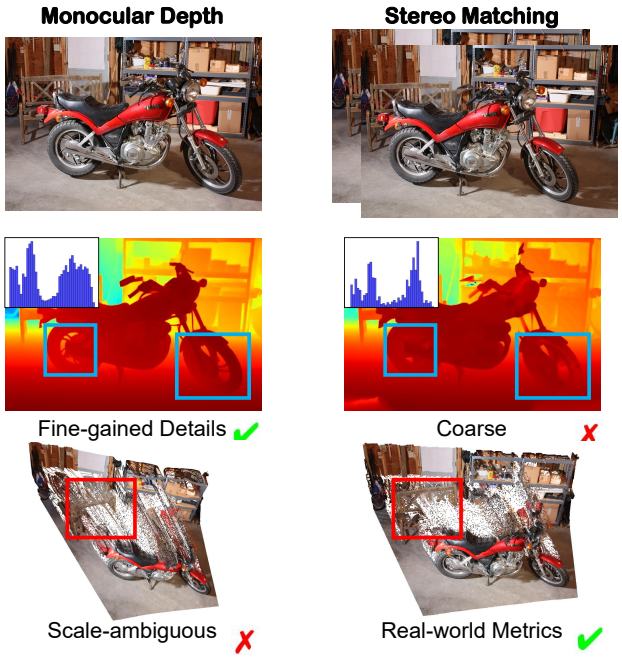


Figure 1: Comparison between Monocular Depth and Stereo Matching. Monocular Depth approach provides fine-grained details but suffers from scale ambiguity, whereas Stereo Matching delivers real-world metrics but with coarser results. We carefully design data generation and dynamic scale- and shift- invariant loss to perform knowledge transfer from monocular to stereo.

1 Introduction

Stereo matching, the task of estimating disparity from two input images, plays an important role in computer vision, powering applications in fields such as robotics [Zhang *et al.*, 2015; Zhang *et al.*, 2024a], autonomous driving [Orb, 2017], and augmented reality [Yang *et al.*, 2019a]. Recent advances in deep learning have led to the development of learning-based methods [Chang and Chen, 2018; Xu *et al.*, 2023; Zhang *et al.*, 2021; Zheng *et al.*, 2024] that achieve state-of-the-art performance. However, the generalization of these methods is highly dependent on the availability of high-quality and diverse training data. Existing datasets can be broadly categorized into synthetic and real-world datasets.

Synthetic datasets, generated from blending engines [Mayer *et al.*, 2016], mainly cover indoor environments. Models trained on these datasets perform well in indoor environments but show reduced effectiveness in outdoor scenarios due to greater visual and contextual variability. In contrast, real-world datasets [Geiger *et al.*, 2012; Menze and Geiger, 2015] typically rely on LiDAR sensors for the ground truth disparity label. However, LiDAR-based ground truth is both costly and time-consuming to acquire and often suffers from issues such as misalignment and incompleteness. As a result, real-world stereo datasets often have sparse and noisy ground truth, which provides insufficient supervisory signals during training, limiting the generalization and accuracy of stereo matching models.

To address these challenges, several approaches have focused on refining network architectures [Xu *et al.*, 2023;

Lipson *et al.*, 2021; Zhang *et al.*, 2025; Zhang *et al.*, 2024b] or generating additional datasets [Liang *et al.*, 2023; Guo *et al.*, 2024; Wang *et al.*, 2024b; Liang and Fu, 2025]. Although these strategies lead to some performance improvements, they still do not resolve the domain gaps between synthetic and real-world data or address the sparsity of labels in real-world datasets well. Nowadays, the rise of large-scale vision foundation models (VFMs) has shown remarkable success and great generalization in a variety of tasks, including monocular depth estimation with models like DepthAnything [Yang *et al.*, 2024] and segmentation with DINOv2 [Oquab *et al.*, 2023]. Despite their success in many domains, the potential of VFMs for stereo matching remains underexplored, mainly because they are designed for monocular tasks and lack explicit geometric constraints.

In this paper, we propose **BooSTer**, a novel framework that combines vision foundation models and large-scale mixed image sources, including synthetic, real-world, and single-view datasets, to **Boo**sting Zero-shot **STer**eo Matching performance. One of the key challenges in stereo matching is the scarcity of labeled data, particularly in real-world datasets. To address this, we introduce a monocular-depth-guided stereo data generation pipeline, which enriches training data from single-view images and allows the model to better generalize across diverse domains. Additionally, we tackle the issue of sparse and incomplete ground truth in real-world stereo datasets by using a monocular depth network to generate pseudo-labels and applying a dynamic scale- and shift-invariant loss to ensure robust learning of relative depth information. In addition, to bridge the domain gap between synthetic and real-world data, we integrate vision foundation model as an encoder to improve feature representations and boost the model's generalization capability. Experimental results demonstrate the effectiveness of our method, showcasing its competitiveness in zero-shot stereo matching.

In summary, our main contributions can be summarized as follows:

- we present a novel framework that transfers knowledge from vision foundation models to stereo matching, boosting both generalization and performance.
- we design a stereo data training framework guided by monocular depth estimation, enabling the creation of large-scale real-world stereo data and introducing a dynamic scale- and shift-invariant loss, to enhance stereo matching quality.
- we employ a hybrid encoder combining a large-scale vision foundation model (DINOv2) and a traditional convolutional feature extractor to obtain global context and local details crucial for stereo matching.

2 Related Work

In this section, we review the most relevant studies on stereo matching, stereo dataset generation, and vision foundation models.

2.1 Stereo Matching Method

With the successful application of learning-based methods [Kendall *et al.*, 2017; Guo *et al.*, 2019; Chen *et al.*, 2023; Zou *et al.*, 2024; Li *et al.*, 2024; Liu *et al.*, 2024; Li and Fu, 2025] across various vision tasks, learning-based stereo matching methods have been replaced traditional optimization methods with CNN networks. GCNet [Kendall *et al.*, 2017] first uses 3D convolutional encoder-decoder architecture to regularize a 4D volume. Following the success of GCNet, PSMNet [Chang and Chen, 2018], GwcNet [Guo *et al.*, 2019] and GANet [Zhang *et al.*, 2019] gradually increase the precision of the network. Besides, cascade methods like CFNet [Shen *et al.*, 2021] is proposed to improve efficiency. RAFT-Stereo [Lipson *et al.*, 2021] proposes to recurrently update disparity map using local cost values retrieved from the all-pairs correlations. IGEV [Xu *et al.*, 2023] constructs a new module to encode non-local geometry and context information. StereoBase [Guo *et al.*, 2023] serves as a strong baseline model in deep stereo-matching by combining all existing architectures. While these methods adopt increasingly complex architectures, few have focused on improving the encoder, which is crucial for stereo matching performance.

2.2 Stereo Datasets

Stereo datasets are typically categorized into synthetic and real-world. Real-world datasets such as KITTI12 [Geiger *et al.*, 2012], KITTI15 [Menze and Geiger, 2015], and ETH 3D [Schops *et al.*, 2017] provide limited pairs with sparse LiDAR-based or indoor depth labels. Larger real-world datasets like DIML [Cho *et al.*, 2021], HRWSI [Xian *et al.*, 2020], and IRS [Wang *et al.*, 2021b] offer more data, but their ground truth (GT) obtained via stereo matching limits its deep stereo performance. Synthetic datasets like SceneFlow [Mayer *et al.*, 2016] use computer graphics (CG) to generate dense GT, but suffer from domain gaps due to limited realism. Generating additional training data [Watson *et al.*, 2020; Wei *et al.*, 2022; Gao *et al.*, 2025] is often used to mitigate this issue. Alternatives such as MfS [Watson *et al.*, 2020], which synthesize stereo from single-view images, often introduce artifacts like holes and collisions, impairing performance.

2.3 Vision Foundation Models

With advancements in computing power and the growth of large-scale datasets, a growing number of vision foundational models (VFMs) [Yang *et al.*, 2024; Radford *et al.*, 2021; Ravi *et al.*, 2024; Oquab *et al.*, 2023] with strong generalization and high performance have emerged. For instance, in the area of monocular depth estimation, models like DepthAnything [Yang *et al.*, 2024] have demonstrated outstanding generalization performance. Moreover, numerous studies have introduced more general backbone networks through self-supervised pre-training. Among these, the DINO family [Oquab *et al.*, 2023; Darcet *et al.*, 2023] has explored self-supervised learning with visual transformers, achieving improvements across a range of downstream tasks. Despite the impressive generalization and zero-shot capabilities of these VFMs, how to effectively leverage them to enhance stereo matching performance remains an open challenge.

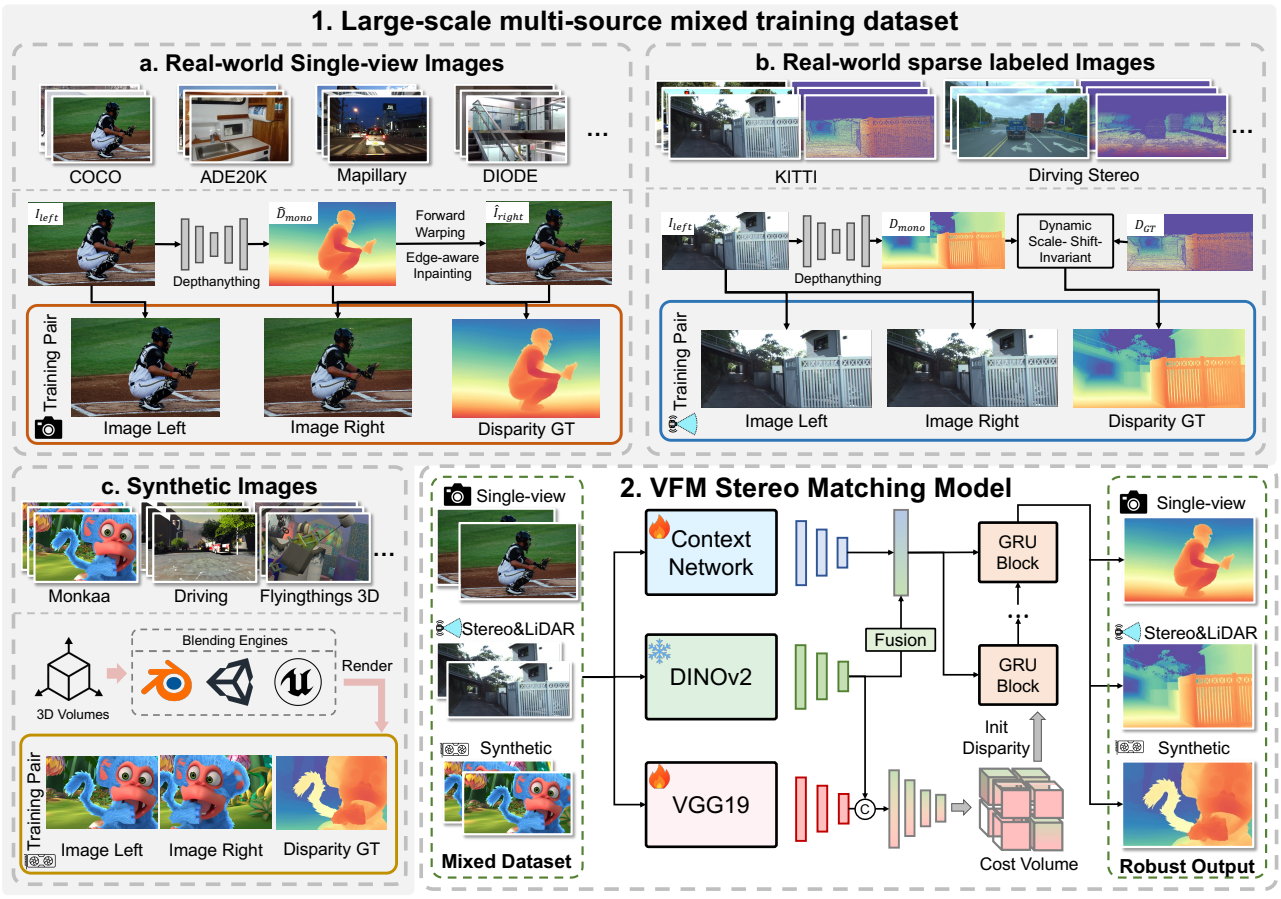


Figure 2: The overall architecture of our proposed method consists of two main parts. 1. **Large-scale multi-source mixed training dataset.** We mix stereo data from various scenarios as a large-scale pre-training dataset. 2. **VFM Stereo Matching Model.** We propose a hybrid encoder structure embedded in VFM to boost the generalization and performance of the algorithm by transferring existing knowledge.

3 Method

In this section, we first introduce the motivation and formulation of **BooSTer**. Then, we present the details of our DFM Guided Training Framework and VFM integrated hybrid encoder architecture.

3.1 Motivation and Formulation

In the stereo matching task, the input consists of a pair of stereo images, $\mathbf{I}_l$ and $\mathbf{I}_r$. The goal is to generate a dense disparity map $\mathbf{D}_{pred}$. To achieve this, a stereo matching model f_θ parameterized by θ is commonly trained as

$$\mathbf{D}_{pred} = f_\theta(\mathbf{I}_l, \mathbf{I}_r). \tag{1}$$

To guide the model's learning, a sparse supervision loss is defined based on the available sparse depth annotations

$$\mathcal{L}_{sup} = \|\mathbf{M}_{val} \odot (\mathbf{D}_{GT} - \mathbf{D}_{pred})\|_1, \tag{2}$$

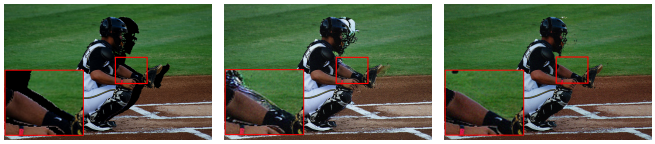
where $\mathbf{D}_{GT}$ denotes the ground truth disparity map, and $\mathbf{M}$ represents the valid mask for the ground truth disparity $\mathbf{D}_{GT}$. $\mathbf{M}_{val}$ is a matrix of valid masks for disparity maps. In practice, the stereo matching model f_θ consists of two components, *i.e.*, a feature extraction encoder f_e, and a disparity refinement module f_d.

The performance and generalization of stereo matching algorithms are often constrained by the limited scale of real-world datasets and the encoder's challenge in learning robust features from synthetic data. While synthetic datasets provide large amounts of data, they fail to capture real-world complexities, such as lighting, texture, and noise, making models trained on them struggle to generalize effectively to real-world environments.

To address these limitations, we aim to improve both generalization and performance by expanding the training dataset and incorporating a hybrid encoder. We introduce a large-scale mixed dataset of real-world images alongside synthetic datasets and integrate vision foundation model into our encoder. This hybrid encoder leverages the pre-trained semantic and feature extraction capabilities of VFM, enabling more accurate and comprehensive feature extraction and enhancing the model's generalization and performance.

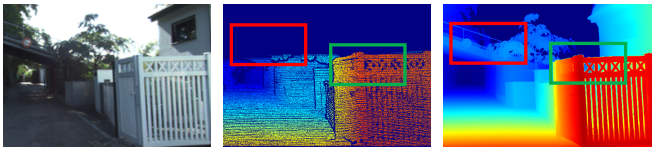
3.2 Learning from Single-view Images

A disparity map is defined as the per-pixel horizontal displacement between the corresponding locations of every pixel from the first view to the second. In this case, from the left image $\mathbf{I}_l$ to the right image $\mathbf{I}_r$. Described in mathematical



(a) Warping Image (b) Inpaint w/o EA (c) Inpaint w/ EA

Figure 3: Comparison between naive inpainting module and our inpainting module. (a) warped right-view image with occulusion holes, (b) image inpainted from (a) using SD, and (c) right image warping with our inpainting module.



(a) RGB Image (b) GT Disparity (c) Mono-Prediction

Figure 4: Comparison between disparity from KITTI and from monocular model. (a) RGB Image, (b) ground-truth from LiDAR, and (c) disparity from monocular estimation.

form, the disparity map can be defined as

$$\mathbf{D}(i) = x_l(i) - x_r(i') \tag{3}$$

where $\mathbf{D}(i)$ is the disparity value of pixel i, and $x_l(i)$ and $x_r(i')$ are the horizontal coordinates of pixel i and i' in the left and right view images, respectively.

Given any disparity map, we can warp every left view pixel i to the right view pixel i' using the disparity map. Therefore, we use the monocular depth estimation model to formulate the disparity map $\mathbf{D}_{mono}$. The disparity map from the monocular model is a relative disparity map ranging from $[0, 1]$. We scale $\mathbf{D}_{mono}$ with a random factor α to convert it into the pixel-wise disparity map $\mathbf{D}'_{mono}$. The transformation can be formulated as

$$\mathbf{D}'_{mono} = \alpha \mathbf{D}_{mono}, \tag{4}$$

where $\alpha \in [d_{min}, d_{max}]$. To boost the diversity of the generated data, we randomly sample the scaling factor $alpha$ from a uniform distribution $U(d_{min}, d_{max})$, where d_{min} and d_{max} are the minimum and maximum scaling factors, respectively.

Edge-Aware Inpainting Module. Random scaling and forward warping address the challenge of generating stereo image pairs. However, the generated right-view images often contain occlusion holes in regions that are invisible in the left-view images. To fill in these missing parts, we use Stable Diffusion (SD). However, as shown in Figure 3, simply using an inpainting model tends to blend the front and back contents. To address this issue, we present an Edge-Aware (EA) inpainting module to mitigate the blending problem. According to Figure 3, this blending issue arises from a lack of object edge information. Therefore, we first generate the object edge mask $\mathbf{M}_e \in [0, 1]^{H \times W}$ from $\mathbf{D}'_{mono}$. We detect the horizontal object edges to create the edge mask using

$$\mathbf{M}_e(i) = \begin{cases} 1, & if\ \nabla_x(\mathbf{D}'_{mono}(i)) > \tau \\ 0, & otherwise \end{cases}, \tag{5}$$

where $\nabla_x(\cdot)$ computes the horizontal gradient of the disparity map, and τ is the threshold for detecting object edges. We select a few background pixels from $\mathbf{M}_e$ and warp them with the foreground to preserve edge information, and then apply inpainting with the SD model to fill occultation holes. The EA module effectively reduces foreground-background fusion, enhancing the realism of the inpainted image.

3.3 Learning from Sparse Images

As shown in Figure 4, pseudo-labels from monocular depth estimation provide accurate details and infer relative depth. However, these scaleless depth estimates are not suitable as direct supervision for stereo models. Therefore, we propose to transfer both monocular depth estimation and stereo matching model output into same scaleless domain to fix the domain gap. As a result, we introduce dynamic scale- and shift- invariant (DSSI) transformation to align the prediction to monocular pseudo depth based on a least-squares criterion

$$(a, b) = \arg\min_{a,b} \sum_{i=1}^{M} \left(a\mathbf{D}_{pred}(i) + b - \mathbf{D}_{mono}(i)\right)^2, \tag{6}$$

where a is scale factor and b is shift factor. Due to the susceptibility of the least squares method to outliers, which compromises the stability of supervision, we conducted outlier detection and removal on the scaled map then recalculated the a and b. The specific mathematical expressions are

$$\mathbf{M}_l(i) = \begin{cases} 1, & if\ (a\mathbf{D}_{pred}(i) + b - \mathbf{D}_{mono}(i))^2 < \tau_l \\ 0, & otherwise \end{cases}, \tag{7}$$

$$\tau_l = Q(a\mathbf{D}_{pred} + b - \mathbf{D}_{mono}), \tag{8}$$

where τ_l is the threshold of masking, and $Q(\cdot)$ is the Quintiles operation. After filtering, we recompute the scale a' and shift b' based on the remaining value and develop DSSI loss using mean-squared error (MSE),

$$\begin{aligned} \mathcal{L}_{DSSI} &= MSE(\hat{\mathbf{D}}_{pred}, \mathbf{D}_{mono}), \\ \hat{\mathbf{D}}_{pred} &= a'\mathbf{D}_{pred} + b'. \end{aligned} \tag{9}$$

Together with the sparse GT loss, the final loss function is formulated as

$$\mathcal{L} = \mathcal{L}_{sparse} + \beta \mathcal{L}_{DSSI}, \tag{10}$$

where β is the DSSI factor of loss functions.

3.4 VFM Based Hybrid Feature Encoder

With the rapid advancement of deep learning techniques, numerous network architectures have been proposed for stereo matching task, each incorporating distinct module designs. Most stereo matching networks can generally be categorized into a two-stage framework, a feature extraction encoder, and a disparity refinement module based on cost volume.

Here, we employ a hybrid feature extraction encoder by integrating DINOv2, a powerful pre-trained feature extraction model, with the traditional convolutional neural network (CNN) based encoder. DINOv2 captures richer and more robust high-dimensional feature representations, including both

Networks	KITTI12			KITTI15			ETH 3D			Middlebury		
	D1 ↓	EPE ↓	>2px ↓	D1 ↓	EPE ↓	>2px ↓	D1 ↓	EPE ↓	>2px ↓	D1 ↓	EPE ↓	>2px ↓
PSMNet'18	30.51	4.68	42.15	32.14	5.98	44.26	8.79	8.26	10.20	22.54	7.93	33.53
CFNet'21	13.63	2.27	20.07	12.08	2.89	19.04	3.07	2.48	4.03	18.41	6.60	23.91
GwcNet'19	23.04	2.76	33.71	25.19	3.57	36.96	6.31	2.59	7.94	22.16	7.28	29.87
COEX'21	12.07	1.79	20.23	11.00	2.48	19.98	3.71	1.74	5.37	18.19	6.82	25.16
FADNet++'21	11.31	1.77	18.02	13.23	2.97	20.90	11.73	8.36	14.26	15.23	4.87	24.08
CasStereo'20	11.85	1.82	18.84	12.06	2.69	20.26	3.83	1.48	5.42	20.61	7.58	27.39
IGEV'23	4.81	0.95	7.84	5.10	1.20	8.90	1.39	0.45	1.88	8.44	2.10	11.77
Selective-IGEV'24	5.73	1.08	5.73	5.63	1.24	9.60	2.45	1.35	3.03	24.81	23.63	29.11
StereoBase'24	4.98	1.02	8.09	5.47	1.20	9.47	1.28	0.28	1.76	8.31	1.71	11.95
Ours	**3.04**	**0.76**	**5.19**	**3.22**	**0.89**	**6.63**	**0.70**	**0.24**	**1.12**	**7.50**	**1.66**	**11.72**

Table 1: Comparison with stereo matching methods. "Ours" indicates pre-training on our mixed dataset. Models are validated on the KITTI12, KITTI15 and ETH3D training sets as cross-domain validation. By default, all supervised methods have trained using SceneFlow.

semantic and high-level information, while CNN focuses on extracting detailed pixel-wise features. This combination enables the model to process the image from both global and local perspectives, thereby enhancing the performance and effectiveness of the encoder.

Formally, given a pair of left and right images $\mathbf{I}_l, \mathbf{I}_r \in \mathbb{R}^{H\times W\times 3}$, we begin by employing the well-established VGG19 [Simonyan and Zisserman, 2015] as the CNN feature extractor. This network generates multi-level pyramid features, producing feature maps at various scales: $f_c^{(i)} \in \mathbb{R}^{C_i\times \frac{H}{i}\times \frac{W}{i}}$, where $i \in \{4, 8, 16\}$ denotes the scaling factor. These multi-scale features offer a rich and layered representation of the input images. To further enhance the feature extraction process, we introduce DINOv2 to generate an additional high-dimensional feature layer, $f_c^{(32)} \in \mathbb{R}^{C_{32}\times \frac{H}{32}\times \frac{W}{32}}$, which captures high-level semantic and contextual information essential for stereo matching.

Finally, the disparity refinement network is designed following the methodology outlined in [Guo *et al.*, 2023], which can boost the depth estimation by refining the initial disparity maps. The combination of advanced feature extraction and disparity refinement improves accuracy and generalization in stereo matching.

4 Experiment

In this section, we first introduce the datasets used for training and evaluation, as well as the details of our implementation. Then, detailed comparisons are conducted under zero-shot settings. Finally, ablation studies are performed to confirm the effectiveness of our proposed main components.

4.1 Datasets and Evaluation Metrics

ETH3D [Schops *et al.*, 2017] is a widely used dataset with grayscale stereo pairs from indoor and outdoor environments, featuring LiDAR ground truth. It includes 27 training and 20 test frames for low-resolution two-view stereo. We use ETH3D as validation set to evaluate our method.

KITTI 2012 and KITTI 2015 [Geiger *et al.*, 2012; Menze and Geiger, 2015] are well-known benchmarks with 200 labeled training pairs and additional test pairs. Ground truth is provided by sparse LiDAR. KITTI15 includes dynamic scenes with semi-automatically generated ground truth. We evaluate the trained models on the training sets of both datasets, following prior work.

Middlebury [Scharstein *et al.*, 2014] contains two sets of 15 stereo image pairs for training and testing, captured from indoor scenes at three resolutions (full, half, and quarter). For evaluations, only the training pairs at half resolution are used to assess cross-domain generalization.

Mixed training datasets consist of three component. The first is synthetic data, such as SceneFlow [Mayer *et al.*, 2016], which includes over 39,000 stereo frames at 960×540 resolution, but has domain gaps with real-world applications. To address this, we create the Diffusion-based Mono for Stereo (**DiffMFS**) dataset by combining single-view datasets like COCO 2017 [Lin *et al.*, 2014], Mapillary Vistas [Neuhold *et al.*, 2017], ADE20K [Zhou *et al.*, 2017], Depth in the Wild [Chen *et al.*, 2016], and DIODE [Vasiljevic *et al.*, 2019], resulting in 597,727 single-view images. Our method then synthesizes stereo data for training. For real-world datasets, we use Driving Stereo [Yang *et al.*, 2019b] and apply our dynamic scale- and shift-invariant loss to handle missing and sparse LiDAR ground truth. In each training epoch, we randomly sample images from the three datasets with a sampling frequency of 5:6:1.

Evaluation Metrics. We report evaluation results on the average End-Point Error (EPE), D1, and >2px metric, which indicates the percentage of stereo disparity outliers. For all metrics, smaller values indicate better model performance.

4.2 Implementation Details

We follow the setup of recent work [Guo *et al.*, 2023] and train our model on 4 NVIDIA 4090 GPUs. For the DFM used in stereo image generation, we adopt Depth Anything V2 [Yang *et al.*, 2024] due to its great generalization. For VFM in the hybrid encoder, we use DINOv2 [Oquab *et al.*, 2023] as its powerful image understanding capabilities and robust features. In the training stage, we utilize our mixed training dataset. We adopt AdamW with a weight decay of 0.05 as the optimizer and clip gradients if their norms exceed 0.1. Our model is pre-trained from scratch for 600K iterations. We set the batch size to 8.

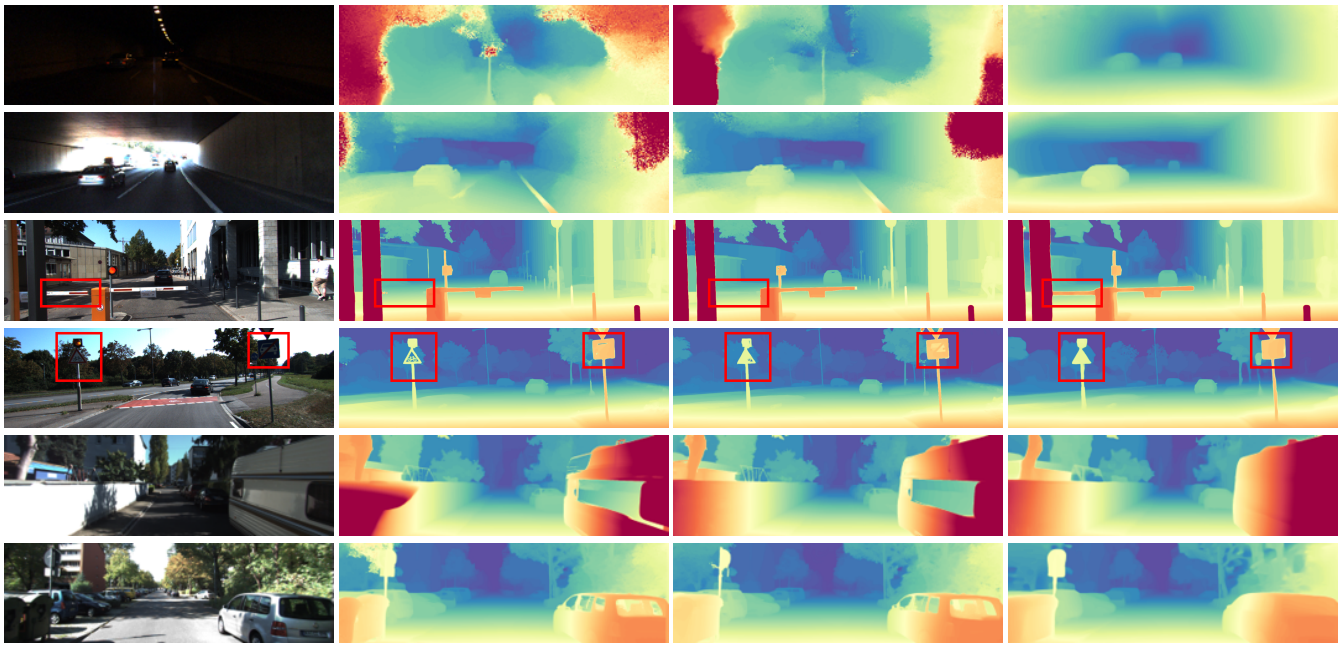


Figure 5: Qualitative results of IGEV and StereoBase trained with Sceneflow and our model trained on our mixed dataset. By default, IGEV and StereoBase are trained using SceneFlow.

4.3 Comparison of Zero-Shot Generalization

To demonstrate the improvement in model generalization, we conducted a cross-domain zero-shot experiment. We selected several state-of-the-art methods for comparison, including PSMNet [Chang and Chen, 2018], CFNet [Shen *et al.*, 2021], GwcNet [Guo *et al.*, 2019], COEX [Bangunharcana *et al.*, 2021], FADNet++ [Wang *et al.*, 2021a], CasStereo [Gu *et al.*, 2020], IGEV [Xu *et al.*, 2023], Selective-IGEV [Wang *et al.*, 2024a], and StereoBase [Guo *et al.*, 2023]. All these methods were evaluated using their best-performing checkpoint, trained on the SceneFlow dataset.

Quantitative Results. As shown in Table 1, we evaluates the effectiveness of our BooSTer. The training pipeline combines simulation data with real-world synthesized data to improve the generalization capabilities of the model. The results demonstrate a substantial improvement in the zero-shot performance, as evidenced by the significant increase in the model's accuracy and robustness across various test sets. Specifically, our method outperforms the baseline model, achieving a notable reduction in disparity estimation errors and providing more precise depth predictions. These results indicate that the mixed training approach not only improves the generalization of the algorithm but also enhances its performance in real-world scenarios, thereby validating the effectiveness of our approach in narrowing the domain gap and improving model robustness.

Qualitative Results. As shown in Figure 5, we present the qualitative results of stereo predictions from models trained on different datasets. In the first two rows, we demonstrate the impact of noise, where the disparity maps generated by StereoBase exhibit noticeable artifacts and inaccuracies, particularly in noisy regions. By incorporating a mixed data training pipeline, our method mitigates the effect of noise, improving disparity estimation and yielding clearer object boundaries. In the next two rows, we highlight the challenges of low-texture regions. StereoBase shows noticeable artifacts around textureless areas, such as poles and signs, while our method progressively refines depth accuracy, especially in regions lacking distinct texture features. In the final two rows, we focus on transparent or reflective materials, where StereoBase produces inaccurate depth maps due to the difficulty in handling such materials. Our method, however, enhances the results, providing more precise depth estimations even in the presence of complex surface reflections and transparency.

4.4 Comparison of Data Generation Methods

To evaluate the effectiveness of our stereo image generation method, we begin by comparing it to the synthetic dataset SceneFlow [Mayer *et al.*, 2016] and the real-world image-based dataset generation method MFS [Watson *et al.*, 2020]. Four distinct model architectures are trained on each dataset using the same training settings.

As shown in Table 2, our method improves generalization across multiple datasets and network architectures, with notable performance gains on the outdoor KITTI datasets. These improvements stem from the enhanced inpainting module, which ensures more coherent content in the generated stereo images, leading to better depth estimation, especially in occlusions and boundary transitions.

4.5 Ablation Study

In this section, we conduct a series of ablation studies to analyze the impact of different module choices in our method.

Networks	Training Dataset	ETH3D			KITTI12			KITTI15		
		EPE ↓	D1 ↓	>2px ↓	EPE ↓	D1 ↓	>2px ↓	EPE ↓	D1 ↓	>2px ↓
PSMNet	SceneFlow	8.272	8.799	10.74	4.683	30.512	43.35	5.986	32.149	44.47
	MFS Dataset	0.534	2.204	3.506	1.009	4.322	6.984	1.605	4.729	8.537
	DiffMFS (Ours)	**0.397**	**1.531**	**2.481**	**0.859**	**4.077**	**6.667**	**1.052**	**4.507**	**8.312**
CFNet	SceneFlow	**0.413**	1.832	2.629	1.010	4.742	7.789	1.839	5.758	9.939
	MFS Dataset	0.712	3.194	4.780	0.922	4.523	7.200	1.113	5.127	9.079
	DiffMFS (Ours)	0.443	**1.434**	**2.500**	**0.881**	**4.169**	**6.669**	**1.030**	**4.711**	**8.562**
IGEV	SceneFlow	0.288	3.610	1.669	1.027	5.135	7.714	1.212	6.034	9.653
	MFS Dataset	0.703	4.657	2.105	0.824	3.546	5.530	1.061	4.912	8.625
	DiffMFS (Ours)	**0.284**	**3.124**	**1.164**	**0.803**	**3.505**	**5.438**	**1.001**	**4.659**	**8.149**
StereoBase	SceneFlow	0.286	1.284	1.762	1.022	4.983	8.091	1.199	5.477	9.478
	MFS Dataset	0.363	1.731	2.343	0.867	4.008	6.571	1.409	4.273	7.874
	DiffMFS (Ours)	**0.250**	**1.047**	**1.392**	**0.777**	**3.468**	**5.475**	**0.984**	**3.904**	**7.176**

Table 2: Zero-shot validation results of models pre-trained on our generated dataset compared with synthetic dataset and data generation method. The results indicates that our stereo data generation method outperforms the state-of-the-art datasets.

Dataset Mixture. We evaluate the effect of different data mixing strategies on the zero-shot performance, testing three strategies: synthetic data, simulation data combined with single-view images, and a mix of simulation, single-view, and real-world images. The results show that stereo images generated from single-view images notably boosts zero-shot performance, suggesting it provides data closer to real-world conditions, helping to bridge the domain gap. Although including real-world data further improves performance, the gain is modest, indicating that its impact is incremental compared to the substantial improvement seen with generated stereo data. This is likely due to the smaller scale and more specific nature of the real-world data used.

Inpainting method. When applying forward warping, holes in the generated image indicate regions with no pixel matched to the single-view image. As shown in Table 3, using an advanced inpainting model outperforms the random hole-filling strategy in [Watson *et al.*, 2020]. However, Stable Diffusion often confuses object and background edges, degrading image quality. In contrast, our Edge-Aware Filling module generates more coherent content, leading to smoother object-background transitions and improved depth estimation.

Stereo Matching Model. To evaluate the performance of the proposed hybrid feature encoder, we applied the same training strategy to the state-of-the-art network, StereoBase [Guo *et al.*, 2023]. When compared to our model, which was trained under identical conditions, our approach demonstrates a effective improvement in zero-shot generalization performance, thereby validating the effectiveness of the encoder optimization proposed in this work.

Exps	Methods	KITTI15	
		EPE ↓	D1 ↓
Data	SF	1.12	5.03
	MONO	0.92	3.78
	SF+MONO	0.90	3.37
	SF+MONO+DS	**0.89**	**3.22**
Inpainting	No	0.99	3.98
	Stable-Diffsuion	0.93	3.56
	Ours	**0.89**	**3.22**
Model	StereoBase	0.91	3.45
	Ours	**0.89**	**3.22**

Table 3: Ablation experiments. Settings used in our method are underlined. "SF" indicates Sceneflow, "MONO" indicates real-world single-view images and "DS" indicates DrivingStereo.

5 Conclusion

In this paper, we propose a novel framework for stereo matching, **BooSTer**, which addresses the challenges of limited labeled data and domain gaps between synthetic and real-world images. By leveraging large-scale mixed image sources, including synthetic, real, and single-view datasets, our approach significantly improves model generalization and performance. We introduce a data generation strategy that combines DFM and diffusion models, enabling the generation of dense stereo matching data from single-view images. Additionally, to overcome the sparse labeled data in real-world datasets, we employ pseudo-mono depth labels and a dynamic scale- and shift-invariant loss function for enhanced supervision. Furthermore, we propose a hybrid encoder structure that integrates VFM, enabling direct knowledge transfer from VFMs at feature level. This approach improves the diversity and generalization of the encoder's features. Through these novel designs,our approach achieves superior performance on real-world zero-shot experiment, scenting an advancement in stereo matching.

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.

Contribution Statement

Yuran Wang and Yingping Liang are equally contributed in this paper. Ying Fu is the Corresponding Author of the paper.

References

[Bangunharcana *et al.*, 2021] 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 *IROS*, pages 3542–3548, 2021.

[Chang and Chen, 2018] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018.

[Chen *et al.*, 2016] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *NeurIPS*, 29, 2016.

[Chen *et al.*, 2023] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *IJCV*, 131(8):2198–2218, 2023.

[Cho *et al.*, 2021] Jaehoon Cho, Dongbo Min, Youngjung Kim, and Kwanghoon Sohn. Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. *arXiv:2110.11590*, 2021.

[Darcet *et al.*, 2023] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

[Gao *et al.*, 2025] 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.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[Gu *et al.*, 2020] 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 *CVPR*, June 2020.

[Guo *et al.*, 2019] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, pages 3273–3282, 2019.

[Guo *et al.*, 2023] 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:2312.00343*, 2023.

[Guo *et al.*, 2024] Wenxuan Guo, Zhiyu Pan, Yingping Liang, Ziheng Xi, Zhicheng Zhong, Jianjiang Feng, and Jie Zhou. Lidar-based person re-identification. In *CVPR*, pages 17437–17447, 2024.

[Kendall *et al.*, 2017] 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 *CVPR*, pages 66–75, 2017.

[Li and Fu, 2025] 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.

[Li *et al.*, 2024] Miaoyu Li, Ying Fu, Tao Zhang, Ji Liu, Dejing Dou, Chenggang Yan, and Yulun Zhang. Latent diffusion enhanced rectangle transformer for hyperspectral image restoration. *TPAMI*, 2024.

[Liang and Fu, 2025] Yingping Liang and Ying Fu. Relation-guided adversarial learning for data-free knowledge transfer. *IJCV*, 133(5):2868–2885, 2025.

[Liang *et al.*, 2023] Yingping Liang, Jiaming Liu, Debing Zhang, and Ying Fu. Mpi-flow: Learning realistic optical flow with multiplane images. In *ICCV*, pages 13857–13868, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[Lipson *et al.*, 2021] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, pages 218–227. IEEE, 2021.

[Liu *et al.*, 2024] 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. *TPAMI*, 2024.

[Mayer *et al.*, 2016] 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 *CVPR*, pages 4040–4048, 2016.

[Menze and Geiger, 2015] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, pages 3061–3070, 2015.

[Neuhold *et al.*, 2017] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 4990–4999, 2017.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

[Orb, 2017] Mur-Artal R Tardós JD Orb. Slam2: an open-source slam system for monocular, stereo, and rgb-d cameras. *T-RO*, pages 1255–1262, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv:2408.00714*, 2024.

[Scharstein *et al.*, 2014] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, pages 31–42, 2014.

[Schops *et al.*, 2017] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, pages 3260–3269, 2017.

[Shen *et al.*, 2021] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *CVPR*, pages 13906–13915, 2021.

[Simonyan and Zisserman, 2015] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, May 2015.

[Vasiljevic *et al.*, 2019] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv:1908.00463*, 2019.

[Wang *et al.*, 2021a] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. Fadnet++: Real-time and accurate disparity estimation with configurable networks. *arXiv:2110.02582*, 2021.

[Wang *et al.*, 2021b] 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.

[Wang *et al.*, 2024a] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *CVPR*, pages 19701–19710, 2024.

[Wang *et al.*, 2024b] Yuran Wang, Yingping Liang, Hesong Li, and Ying Fu. Mono2stereo: Monocular knowledge transfer for enhanced stereo matching. *arXiv:2411.09151*, 2024.

[Watson *et al.*, 2020] Jamie Watson, Oisin Mac Aodha, Daniyar Turmukhambetov, Gabriel J Brostow, and Michael Firman. Learning stereo from single images. In *ECCV*, pages 722–740, 2020.

[Wei *et al.*, 2022] Kaixuan Wei, Ying Fu, Yinqiang Zheng, and Jiaolong Yang. Physics-based noise modeling for extreme low-light photography. *TPAMI*, pages 8520–8537, 2022.

[Xian *et al.*, 2020] Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *CVPR*, pages 611–620, 2020.

[Xu *et al.*, 2023] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *CVPR*, pages 21919–21928, 2023.

[Yang *et al.*, 2019a] 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 matching algorithms. *IEEE Internet of Things Journal*, 7(4):2521–2530, 2019.

[Yang *et al.*, 2019b] 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 *CVPR*, pages 899–908, 2019.

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

[Zhang *et al.*, 2015] 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.

[Zhang *et al.*, 2019] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, pages 185–194, 2019.

[Zhang *et al.*, 2021] Fan Zhang, Yu Li, Shaodi You, and Ying Fu. Learning temporal consistency for low light video enhancement from single images. In *CVPR*, pages 4967–4976, 2021.

[Zhang *et al.*, 2024a] Fan Zhang, Shaodi You, Yu Li, and Ying Fu. Atlantis: Enabling underwater depth estimation with stable diffusion. In *CVPR*, pages 11852–11861, 2024.

[Zhang *et al.*, 2024b] 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.

[Zhang *et al.*, 2025] Yingkai Zhang, Zeqiang Lai, Tao Zhang, Ying Fu, and Chenghu Zhou. Unaligned rgb guided hyperspectral image super-resolution with spatial-spectral concordance. *arXiv:2505.02109*, 2025.

[Zheng *et al.*, 2024] 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.

[Zhou *et al.*, 2017] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.

[Zou *et al.*, 2024] Yunhao Zou, Ying Fu, Tsuyoshi Takatani, and Yinqiang Zheng. Eventhdr: From event to high-speed hdr videos and beyond. *TPAMI*, 2024.