Structure-grounded Training Strategies Aid Generalization in Stereo Matching

Liangxun Ou, Yuhui Liu, Zhenyang Li, Xiaoyang Bai, Yifan Peng*
The University of Hong Kong, Hong Kong SAR
{liangxun.ou, liuyuhui, lizy23}@connect.hku.hk, {xybai, evanpeng}@hku.hk,

Abstract

*Stereo matching networks can suffer from generalization challenges when trained on synthetic data and deployed in real-world settings. While existing methods rely on fine-tuning or pre-trained vision foundation models for cross-domain robustness, we revisit this gap from a training perspective and explore a structure-grounded training design that directly improves generalization of RNN-based stereo matching models using only a limited amount of synthetic stereo data, without changing the network architecture or adding any inference overhead. Specifically, we target all three main modules of a typical stereo matching pipeline: in **cost volume construction**, we enhance geometric cues through data augmentation; in **context encoding**, we strengthen semantic guidance via auxiliary multi-task context supervision; in **recurrent disparity refinement**, we regulate update dynamics with depth-update regularization. Experiments on multiple mainstream architectures and diverse real-world datasets suggest consistent gains in robustness, improving RAFT-Stereo by 6.6% on KITTI 2015, IGEV-Stereo by 13.7% on Middlebury, and DLNR by 55.4% on ETH3D. These insights reveal the previously overlooked importance of structure-grounded training design for achieving reliable stereo depth estimation under data-scarce, domain-shifted conditions.*

1. Introduction

Stereo matching estimates per-pixel metric depth from rectified image pairs by computing disparity, the horizontal shift of a 3D point between the two camera views, whose magnitude is inversely related to distance and thus serves as a direct proxy for depth. It supports applications in autonomous driving, robotics, and augmented reality [9, 33]. Over the past few years, stereo matching architectures have shifted from 3D-CNN cost volume regression [11, 38] to RAFT-inspired iterative refinement [32]. However, much progress remains focused on benchmarks: models pretrained on synthetic datasets [22] and fine-tuned on real-world datasets (KITTI [10], Middlebury [29], ETH3D [30]) often overfit dataset statistics and degrade in uncurated scenes [14, 24]. Although this approach achieves strong performance on standard benchmarks, it biases optimization toward the narrow statistics of those datasets. Once deployed in uncurated environments, many state-of-the-art models exhibit pronounced error spikes [3, 16, 27]. A recent trend grafts pretrained vision foundation models (VFMs) into stereo pipelines to supply powerful semantic priors [7, 36]. However, integrating VFMs increase parameters and latency in exchange for improved robustness, limiting real-time deployment of these methods.

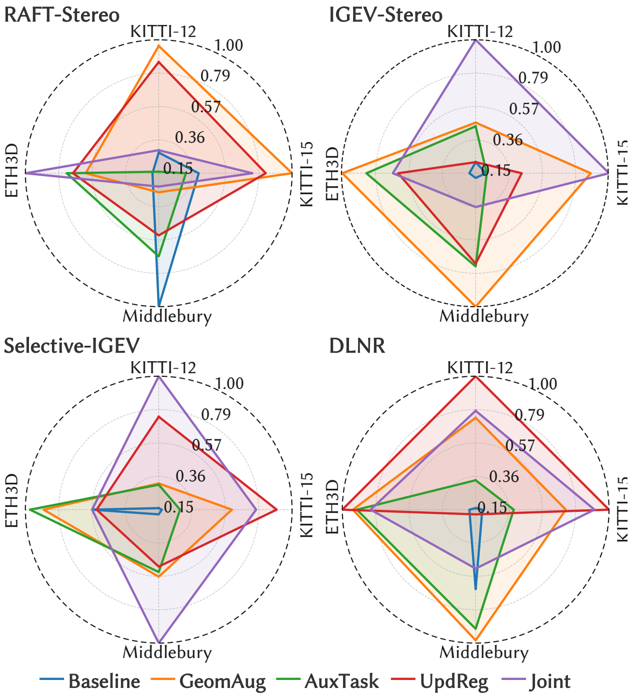

Figure 1. Radar charts comparing four mainstream stereo matching frameworks across four datasets (KITTI 2012, KITTI 2015, Middlebury, ETH3D) under our training strategies (**GeomAug**, **AuxTask**, **UpdReg**, and **Joint**). We use a within-dataset, within-model normalized metric combining $<$ 3 px (%) and EPE, where larger polygons indicate higher depth accuracy. See Sec. 5 for the evaluated strategies and metrics, and the Supplementary Material for normalization details.

[*] Corresponding author.

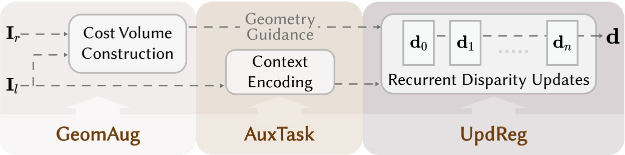

Figure 2. Overview of an RNN-based stereo matching pipeline. Each of our investigated three strategies targets a specific module in this pipeline, as detailed in Sec. 4.

To prevent models from overfitting and learning dataset-specific rather than generalizable knowledge, we investigate mainstream stereo matching methods in light of their pipeline structures, which are typically comprised of: (i) *cost volume construction*, (ii) *context encoding*, and (iii) *recurrent disparity updates* [18, 39], as illustrated in Fig. 2. We find that each module exhibits limitations. Firstly, cost volume is susceptible to matching ambiguity caused by geometric factors, including occlusions, low texture, and repetitive patterns, which compromises matching accuracy [1]. Secondly, context encoding only provides weak priors when faced with domain shift, hindering its ability to guide matching in unseen data. Thirdly, recurrent disparity updates may converge early to overly narrow disparity ranges, restricting the model's capacity to recover challenging depth variations.

These observations motivate a training paradigm that improves real-world generalization from limited synthetic data, rather than merely optimizing performance within a single target domain. Instead of redesigning the backbone model, we explore three lightweight yet synergistic strategies for the three modules respectively to align them better with their intrinsic objectives: **Geometry-oriented data augmentation.** We inject "Ribbon" and "Blob" perturbations into stereo image pairs, enriching synthetic data with regions where stereo correspondence becomes unreliable due to weakened geometric cues or fine structural complexity. This makes *cost volume construction* more robust to the geometric challenges it often faces. **Auxiliary multi-task context supervision.** We repurpose *encoding context* to predict semantic segmentation and edge masks. This leverages the encoder's role as a source of global priors, making its features more object-aware and boundary-sensitive, which helps in ambiguous regions. **Depth-update regularization.** We design a simple regularization loss that discourages *recurrent disparity update* from stopping too early, pushing it to explore wider disparity ranges and to better handle complex depth variations.

The aforementioned strategies are architecture-agnostic without incurring extra inference cost. Across multiple RNN-based baselines, these strategies consistently narrow the sim-to-real gap on diverse real-world benchmarks, as illustrated in Fig. 1, indicating that carefully designed training strategies can unlock robustness under data scarcity. In summary, our main contributions are threefold:

- We investigate a module-level diagnosis of how the cost volume, context branch, and recurrent updater in RNN-based stereo matching models behave under domain-shifted conditions.
- We explore three *plug-and-play* training strategies to improve model generalizability: geometry-oriented augmentation, auxiliary multi-task context supervision, and depth-update regularization, with no pipeline change or inference overhead.
- We demonstrate strong cross-dataset generalization on real-world benchmarks across multiple RNN-based baselines trained only on synthetic data.

2. Related Work

Stereo Matching Methods. Deep learning has become the dominant paradigm for stereo depth estimation, with networks learning to extract and compare features across rectified image pairs [20, 38]. Early approaches such as PSMNet [4] built 4D cost volumes and regularized them with 3D CNNs, which yielded strong benchmark accuracy but incurred high computational and memory costs [40]. To scale better, later methods estimate an initial disparity and iteratively refine it from compact correlation features [18, 32]. RAFT-Stereo [32] exemplifies this shift with recurrent units that propagate matching cues across iterations, while subsequent variants streamline computation and focus refinement on uncertain regions [35].

Despite steady architectural progress [41], stereo matchers still struggle in real-world scenes with limited texture, repeated patterns, occlusions, or complex geometry. Local warping and restricted receptive fields exacerbate depth ambiguity, leading to structural distortions and loss of fine detail. These limitations motivate the need for stronger global reasoning and improved robustness across domains.

Robustness in Stereo Matching. To mitigate these failures, many works introduce semantic priors beyond photometric matching. Early approaches rely on object-level cues such as edges or segmentation masks [31, 44], but their coarse granularity limits performance on thin or textureless structures. Monocular depth has also been used as a dense structural prior [19, 45], at the cost of scale ambiguity and prior-induced bias. Beyond explicit priors, several methods improve cross-domain robustness through representation and training regularization, including hierarchical visual transformations that suppress shortcut learning [5], feature-consistency constraints that stabilize cross-view matching [46], and masked representation learning with self-supervised reconstruction to enhance structure-aware generalization [28].

An alternative line of work departs from the rectified stereo assumption altogether. Feedforward geometry models such as DUST3R [34] and its variants (e.g., MASt3R [17]) directly predict cross-view geometry from unrectified image pairs in a single forward pass which exhibit strong cross-view generalization. However, their accuracy on standard rectified stereo benchmarks remains below that of specialized stereo matchers.

More recently, vision foundation models (VFMs) have been incorporated into stereo pipelines to improve robustness. Approaches that integrate foundation model features enable stereo matching with strong generalization and competitive accuracy across datasets [2, 7, 13, 36]. But these gains come with increased computational cost, which limits their practicality in resource constrained settings [12].

Generalization is further shaped by data and training protocols. Stereo models are typically pre-trained on synthetic datasets and fine-tuned on limited real data, where domain gaps in appearance and geometry persist and common augmentations provide limited robustness [22, 26, 42]. These observations motivate training paradigms that improve robustness under simulated supervision by aligning training signals with internal modules, rather than introducing heavier external components. Our work follows this pathway.

3. Structure-grounded Diagnosis

3.1. Preliminary

Given rectified image pair $(\mathbf{I}_l, \mathbf{I}_r)$, mainstream RNN-based stereo matching frameworks follow a *three-stage* pipeline:

3D cost volume construction. Multi-scale feature maps $\mathbf{F}_l$ and $\mathbf{F}_r$ are extracted, and the cost volume $\mathcal{V}_\eta \in \mathbb{R}^{H\times W\times D}$ is constructed by computing the matching cost for each pixel x and disparity d as $\mathcal{V}_\eta(x,d) = \text{sim}\big(\mathbf{F}_l(x), \mathbf{F}_r(x-d)\big)$, where $\text{sim}(\cdot,\cdot)$ denotes a similarity function. This cost volume forms the geometric foundation for subsequent recurrent disparity refinement.

Context encoding. The context encoder $\mathcal{F}_\theta$ extracts spatial and semantic cues from the left image to initialize the hidden state for recurrent disparity refinement. In most frameworks, the context encoder follows the design of RAFT's shallow residual network, which encodes scene layout and object-level cues [20, 32].

Disparity refinement. The iterative refiner $\mathcal{H}_\psi$ commonly uses convolutional gated recurrent units to fuse cost volume evidence with contextual cues and correct the disparity field. The final disparity is:

$$\mathbf{d}_N = \mathbf{d}_0 + \sum_{n=1}^{N} \Delta\mathbf{d}_n, \tag{1}$$

$$\text{where } \Delta\mathbf{d}_n = \mathcal{H}_\psi\big(\mathbb{L}\{\mathcal{V}_\eta(\mathbf{I}_l, \mathbf{I}_r), \mathbf{d}_{n-1}\},\ \mathcal{F}_\theta(\mathbf{I}_l)\big). \tag{2}$$

Here $\mathcal{V}_\eta(\mathbf{I}_l, \mathbf{I}_r)$ denotes the cost volume built by method η (e.g., correlation [22], concatenation [15]), $\mathbb{L}$ samples the cost volume at $\mathbf{d}_{n-1}$ to retrieve similarity. The initialization $\mathbf{d}_0$ is set zero [20] or produced by a 3D-CNN estimator [41]; after N iterations we output the final disparity $\mathbf{d}_N$.

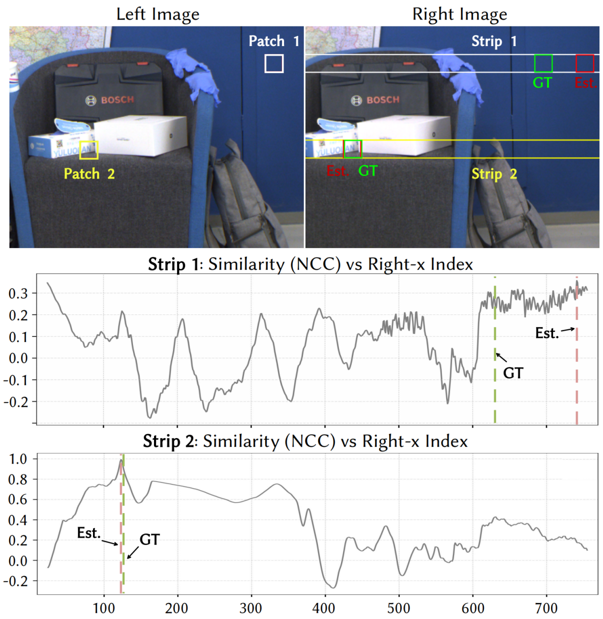

Figure 3. Cost volume intuition and failure cases. Top: two left-image patches and their 1D search strips in the right image. Bottom: normalized cross-correlation (NCC) profiles along each strip (green dashed: GT, red dashed: estimated peak). Strip 1 (low texture) exhibits flat or multi-modal responses, leading to ambiguous matches and deviation from GT, whereas Strip 2 (rich texture) shows a single sharp peak at GT, yielding a reliable match.

3.2. Analysis and Motivation

Cost volume robustness. Ideally, a cost volume exhibits a single, sharp peak at the ground-truth disparity, reflecting accurate geometric correspondence. As illustrated in Fig. 3, real-world effects such as texture sparsity, repetitive patterns, and view-dependent appearance often flatten or multi-modulate the similarity profile, obscuring the true match. These distortions degrade the cost-volume landscape and hinder downstream refinement, frequently resulting in persistent residual errors.

Such challenges are not well addressed by existing training pipelines [18, 35, 39], whose augmentation strategies are largely adapted from semantic vision tasks such as object detection. Geometric augmentations (for example, random cropping, rotation, or small spatial shifts) modify pixel arrangements without altering their underlying values [8], while photometric augmentations (such as changes in saturation, gamma, or contrast) adjust pixel intensities to simulate appearance variations [25]. Both families operate only at the image level and do not diversify geometric or disparity cues, leaving models vulnerable to stereo-specific failure modes. To this end, we present a **geometry-oriented data augmentation** strategy that perturbs disparity maps to emulate structurally challenging scenarios in stereo matching. It comprises: (1) *Ribbon Injection*, which overlays thin, stochastic depth ridges to induce fine-scale ambiguity and

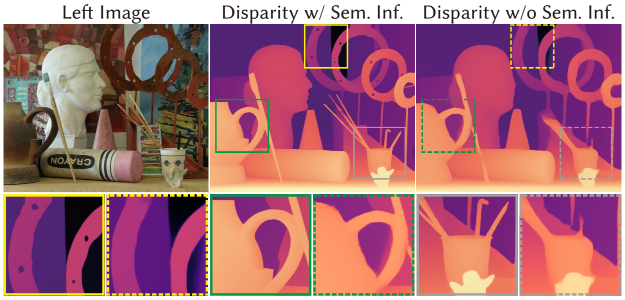

Figure 4. Disparity estimation results of RAFT-Stereo on a Middlebury example with and without semantic information. With the output of the context encoder masked out, the model fails to recover fine object details such as holes and thin rods.

local discontinuities; and (2) *Blob Injection*, which inserts randomly shaped, hole-free depth patches to mimic coherent structures such as occlusions or protrusions. These augmentations are applied to images from the original datasets to create a new dataset with more complex geometry while maintaining stereo consistency. They are plug-and-play yet task-specific, fostering feature representations that are robust to real-world geometric irregularities.

Semantic regularization. In real-world scenes, stereo matching often fails due to local inconsistency, where noise or limited receptive fields disrupt local structures, and stereo inconsistency, where viewpoint changes alter object positions and break correspondence. In both cases, leveraging semantic cues can improve object-level consistency, sharpen boundaries, and reduce artifacts (Fig. 4). Prior works such as SegStereo [43] and SSPCV-Net [37] introduce dedicated semantic networks with explicit semantic supervision, and inject the resulting semantic features into cost volume construction to guide matching, primarily in earlier stereo architectures. To extend this semantic enhancement principle to more modern stereo pipelines, we adopt this semantic enhancement principle through **auxiliary multi-task context supervision**: during training, the context network jointly predicts *object index maps* and *edge maps*, which are optimized against ground-truths. This approach injects region-level coherence and boundary awareness to resolve ambiguities and preserve global geometry without altering the original network architecture or injecting additional semantic features into the matching pipeline.

Disparity refinement dynamics. RNN-based disparity refinement commonly exhibits a rapid decay in update magnitude, yet convergence often stalls before reaching the correct disparity, indicating premature settling into local optima, as shown in Fig. 5. By introducing **depth-update regularization** during training, the refinement process maintains effective updates and yields consistent performance gains on the test set (Sec. 5), demonstrating improved generalization across varying disparity distributions.

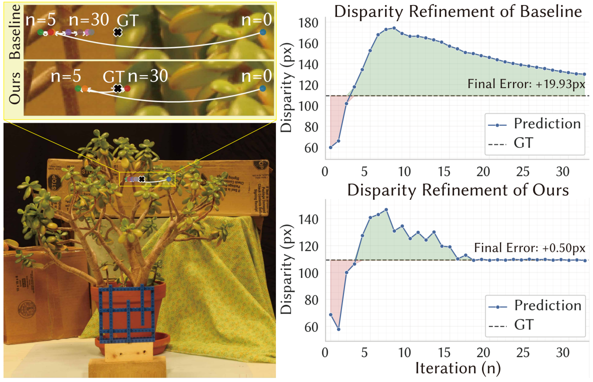

Figure 5. Disparity refinement dynamics of RAFT-Stereo. *Left*: left-image crops with refinement trajectories; colored markers denote predicted matches every 5 iterations, and the dark cross indicates the ground truth (GT). The top row shows the baseline model, while the bottom row shows the same architecture trained with **depth-update regularization**. *Right*: corresponding disparity trajectories over iterations.

4. Approach

4.1. Geometry-oriented Data Augmentation

To enrich stereo pairs with complex geometry, we overwrite the left disparity map with structured masks (Ribbons and Blobs) and synthesize the stereo-consistent right image from the modified disparity and the original left image.

Mask generation on the pixel grid. We first generate binary masks on a coarse pixel grid $\Omega_b = \{(u, v) \in \mathbb{Z}^2 : 0 \leq u < W, 0 \leq v < H\}$, which are subsequently placed onto the target disparity map. Two mask modes are employed, as shown in Fig. 6: **(a) Ribbon**. Thin continuous ridges are produced by sampling multiple random curves. For each curve k, we sample m_k control points $\mathcal{P}_k \subset \Omega_b$ uniformly and order them left-to-right. A quadratic B-spline γ_k is then interpolated through $\mathcal{P}_k$, producing a polynomial curve $\mathcal{C}_k$, which is then rasterized with a random stroke thickness t_k. We obtain the ribbon mask with n_c curves as:

$$\mathbf{M}_b^{\text{rib}}(u, v) \;=\; \min\left(1, \sum_{k=1}^{n_c} \mathbf{1}\big[\,\text{dist}\left((u, v), \mathcal{C}_k\right) \leq \tfrac{t_k}{2}\big]\right), \tag{3}$$

where $\mathbf{1}[\cdot]$ is the indicator function and $\text{dist}(\cdot, \mathcal{C}_k)$ denotes the Euclidean distance to $\mathcal{C}_k$. **(b) Blob**. Coherent occluding shapes are commonly formed as filled, hole-free polygonal blobs. For each blob i we sample a center c_i and number of vertices v_i. Each vertex j is defined in polar coordinates relative to c_i by a radius $r_{i,j}$ and an angle $\theta_{i,j}$. Angles are sorted to avoid self-intersection. These parameters define a simple polygon Π_i, and the blob mask with n_b polygons is:

$$\mathbf{M}_b^{\text{blob}}(u, v) = \min\left(1, \sum_{i=1}^{n_b} \mathbf{1}\big[(u, v) \in \text{int}(\Pi_i)\big]\right), \tag{4}$$

where $\text{int}(\Pi_i)$ denotes the polygonal interior on Ω_b.

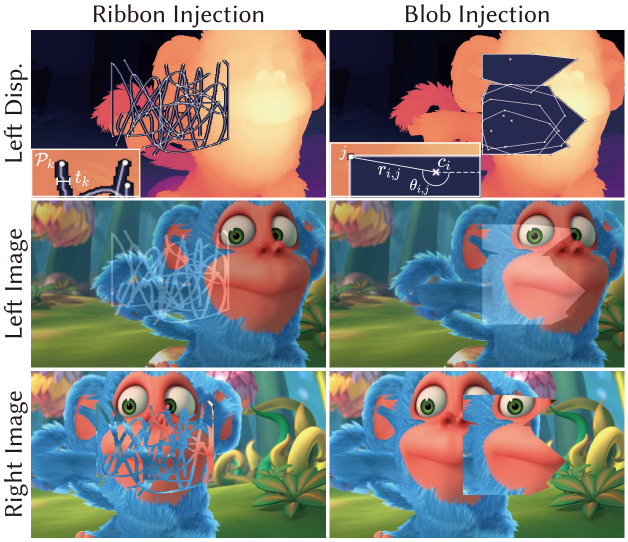

Figure 6. Examples of Ribbon (*left*) and Blob (*right*) injection. For each type, we show the left disparity map with generated mask (*top*), the left image with semi-transparent mask overlay (*mid*), and the synthesized right image (*bottom*). Additional cases are provided in the Supplementary Material.

The generated base-grid mask (Eqs. 3–4) is assigned a constant horizontal disparity offset δ, randomly drawn from integer values in the depth range $[-d_{\max}, d_{\max}]$. This ensures that the bounding box remains within the image and that all pixels in the mask share the same disparity.

Disparity perturbation. Then, we alter the left disparity $\mathbf{d}_l$ within the mask $\mathbf{M}_l$ by replacing the disparity values with the offset δ:

$$\mathbf{d}'_l = \mathbf{d}_l \odot (1 - \mathbf{M}_l) + \delta\, \mathbf{M}_l, \tag{5}$$

where $\odot$ denotes element-wise multiplication.

Right-image synthesis. Finally, a new right image $\tilde{\mathbf{I}}_r$ is produced by backward warping the left image $\mathbf{I}_l$ using the updated $\mathbf{d}'_l$ for all pixels (u, v) within the mask $\mathbf{M}_l$:

$$\tilde{\mathbf{I}}_r(u, v) = \mathbf{I}_l\left(u - \mathbf{d}'_l(u, v),\, v\right). \tag{6}$$

Outside the mask, the original right image is preserved. We keep the left image unchanged. For a feasibility analysis of this approach from the perspective of information entropy. **Please refer to the Supplementary Material for details.**

4.2. Auxiliary Multi-task Context Supervision

To integrate semantic and structural information within the originally encoded context features, we design two lightweight decoders on the context encoder output feature $\mathbf{C}_{\{4,8,16\}} = \{\mathbf{C}_4, \mathbf{C}_8, \mathbf{C}_{16}\}$ at resolution downsampling ratios $1/4$, $1/8$, and $1/16$. These decoders generate auxiliary predictions for *object index maps* and *edge maps* (Fig. 7), which are jointly optimized with disparity estimation. We detail the design of each decoder as follows.

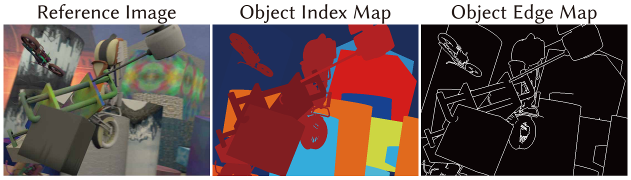

Figure 7. Auxiliary supervision targets. Object index map provides semantic separation, while object edge map highlights geometric boundaries.

Object index prediction. The designed object index decoder $\mathcal{F}_{\text{obj}}$ employs an Atrous Spatial Pyramid Pooling (ASPP) [6] block on $\mathbf{C}_{16}$ to enlarge the receptive field, followed by two upsampling stages with skip connections from $\mathbf{C}_8$ and $\mathbf{C}_4$. Finally, a 1×1 convolution block produces the object-level logits:

$$\hat{\mathbf{O}} = \mathcal{F}_{\text{obj}}(\mathbf{C}_{\{4,8,16\}}) \in \mathbb{R}^{H \times W \times K}, \tag{7}$$

Ground-truth indices $\tilde{\mathbf{O}} \in \{-1, 0, \ldots, K\}^{H \times W}$ are obtained by remapping raw instance IDs into consecutive labels per image, where -1 denotes ignore, 0 denotes background, and K is the maximum foreground index in that image. The loss combines pixel-wise cross entropy with a multi-class Dice term computed over only valid indices and restricted to the active label set of each image:

$$\mathcal{L}_{\text{obj}} = \lambda_{\text{ce}}\, \mathcal{L}_{\text{CE}}(\hat{\mathbf{O}}, \tilde{\mathbf{O}}) + \lambda_{\text{dice}}\, \mathcal{L}_{\text{Dice}}\Big(\text{softmax}(\hat{\mathbf{O}}),\ \tilde{\mathbf{O}}\Big), \tag{8}$$

Edge detection. We also feed the features $\mathbf{C}_{4,8,16}$ into an edge decoder $\mathcal{F}_{\text{edge}}$. Each feature map is upsampled to a uniform $1/4$ resolution, followed by a 1×1 convolution to produce the binary edge prediction:

$$\hat{\mathbf{E}} = \mathcal{F}_{edge}(\mathbf{C}_{\{4,8,16\}}) \in \mathbb{R}^{H \times W}. \tag{9}$$

Given the ground truth edge $\tilde{\mathbf{E}} \in \{0, 1\}^{H \times W}$, we use a combination of binary cross entropy ($\mathcal{L}_{\text{bce}}$) and Dice loss:

$$\mathcal{L}_{\text{edge}} = \lambda_{\text{bce}}\, \mathcal{L}_{\text{bce}}(\hat{\mathbf{E}}, \tilde{\mathbf{E}}) + \lambda_{\text{dice}}\, \mathcal{L}_{\text{Dice}}(\sigma(\hat{\mathbf{E}}), \tilde{\mathbf{E}}), \tag{10}$$

where $\sigma(\cdot)$ is the sigmoid function.

Together, these cues enable the encoded features to represent view-consistent scene structures with meaningful semantics , which improves cost volume robustness and disparity refinement.

4.3. Depth-update Regularization

Recurrent stereo pipelines refine the disparity over N iterative update steps $\{\mathbf{d}_i\}_{i=1}^{N}$. Following [39], we supervise all iterations using a sequence loss:

$$\mathcal{L}_{\text{seq}} = \text{Smooth}_{L_1}(\mathbf{d}_0 - \mathbf{d}_{gt}) + \sum_{i=1}^{N} \gamma^{N-i} \|\mathbf{d}_i - \mathbf{d}_{gt}\|_1, \tag{11}$$

where $\gamma \in (0, 1)$ downweighs earlier predictions.

To encourage more active refinement, we introduce a *depth-update regularization* term:

$$\mathcal{L}_{\text{var}} = -\sum_{i=1}^{N} \gamma^{N-i} \|\mathbf{d}_i - \mathbf{d}_{i-1}\|_1, \tag{12}$$

which encourages each recurrent step to make meaningful progress rather than converging too cautiously. As illustrated in Fig. 5, the baseline updater shows rapid early progress but then stagnates before reaching the ground truth. In contrast, our loss sustains effective updates, guiding the predictions closer to the correct disparity and achieving earlier stabilization.

The final stereo training objective is:

$$\mathcal{L}_{\text{disp}} = \mathcal{L}_{\text{seq}} + \lambda_{\text{var}}\mathcal{L}_{\text{var}} + \lambda_{\text{aux}}(\mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{edge}}), \tag{13}$$

where λ_{var} and λ_{aux} control the strength of the regularization and supervision.

5. Experiments

5.1. Implementation Details

Synthetic training dataset. All models are trained exclusively on the synthetic **Scene Flow** [22] dataset, which is a large-scale synthetic dataset consisting of three subsets: FlyingThings3D, Monkaa, and Driving. Each stereo pair has a resolution of 960×540, with ground truth disparities provided in sharp version and blurred version (with defocus and motion blur). We utilize 35,454 training pairs from all three subsets in the blurred version. The motion boundary maps and segmentations at both object and material levels are provided, enabling auxiliary supervision.

Real-world evaluation datasets. For cross-domain evaluation, we evaluate our models without fine-tuning on four widely adopted benchmarks: **KITTI 2012** [9] contains 194 training and 195 test stereo pairs from real driving scenes, and **KITTI 2015** [23] contains 200 training and additional 200 test pairs, collected under similar driving conditions. For both datasets, sparse ground truth disparities are obtained from LiDAR scans, mainly covering static background regions. **ETH3D** [30] offers 27 training and 20 test stereo pairs, covering both indoor and outdoor scenes with accurate ground truth disparities. **Middlebury** [29] consists of 15 training and 15 test high-resolution indoor stereo pairs, captured under varied lighting and exposure conditions. Its high resolution and photometric diversity make it particularly challenging for models trained on low-resolution synthetic data.

We evaluate four representative recurrent stereo-matching networks on the dataset's training split with ground-truth annotations.

Table 1. Ablation on *geometry-oriented data augmentation*. **Bold** = best, underline = second best.

Params	KITTI-12		KITTI-15		Middlebury		ETH3D	
$(\rho_{\text{aug}}, \gamma_{\text{blob}})$	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)
0.1, 0.5	**0.884**	**95.6**	**1.088**	**94.7**	1.165	93.1	**0.291**	**99.0**
0.1, 0.3	0.935	95.1	1.203	94.3	1.314	91.8	0.419	98.2
0.2, 0.5	0.908	95.4	1.116	94.3	**1.066**	**94.2**	0.329	98.7

Table 2. Ablation on *auxiliary multi-task context supervision*.

Params	KITTI-12		KITTI-15		Middlebury		ETH3D	
$(\lambda_{\text{obj}}, \lambda_{\text{edge}})$	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)
0.1, 0.1	0.948	95.0	1.145	94.2	1.241	94.0	0.302	**99.0**
0.2, 0.2	0.934	95.3	1.152	94.2	**1.128**	**94.1**	**0.285**	**99.0**
0.3, 0.3	0.936	95.2	1.188	94.0	1.144	93.5	0.300	98.8
0.2, 0	**0.882**	**95.6**	**1.099**	**94.7**	1.132	93.7	0.297	98.7
0, 0.2	0.935	95.1	1.133	94.3	1.242	93.4	0.289	98.9

Table 3. Ablation on *depth-update regularization*.

Params	KITTI-12		KITTI-15		Middlebury		ETH3D	
(λ_{var})	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)	EPE (px)	< 3 px (%)
0.1	**0.879**	**95.5**	**1.109**	**94.6**	1.134	93.7	**0.280**	**98.9**
0.2	0.927	95.4	1.135	94.4	**1.129**	**93.8**	0.312	98.8
0.3	1.028	94.5	1.235	93.6	1.448	92.6	0.384	98.4

RAFT-Stereo [20] is the first widely adopted RNN-based stereo method, adapting RAFT [32]'s iterative refinement to disparity estimation via multi-scale convolutional GRUs, enabling accurate and controllable updates.

IGEV-Stereo [39] extends RAFT-Stereo with an enhanced cost volume representation, the Combined Geometry Encoding Volume (CGEV), fusing global geometry cues with local correlations for more reliable matching.

DLNR [47] targets high-frequency detail preservation using a Multiscale Decouple LSTM instead of GRU to maintain fine structures across iterations. A Normalization Refinement module and a Channel-Attention Transformer feature extractor are designed for better cross-domain robustness.

Selective-IGEV [35] builds on IGEV-Stereo with frequency-aware refinement via a Selective Recurrent Unit (SRU) and Contextual Spatial Attention (CSA), improving edge sharpness and textureless-region accuracy.

In our experiments, all models are trained on an NVIDIA RTX 4090 GPU using the AdamW optimizer [21]. Scene Flow images are randomly cropped to 320×736 and trained for 200,000 steps with a one-cycle learning rate schedule (initial learning rate 1×10^{-4}) and a batch size of 3. We apply identical data augmentation, including saturation adjustment, random stretching of images and disparities to simulate diverse disparity distributions, and gamma correction as original papers. Our training retains these augmentations while incorporating the proposed method. All models use 16 update iterations during training and 32 during testing.

5.2. Ablations

We conduct a series of ablation experiments in RAFT-Stereo. Throughout all experiments, performance is assessed using End-Point Error (EPE), which measures the average disparity error in pixels and reflects overall accu-

Table 4. Model comparison w/ our training strategies across datasets. **Bold** = best, underlined = second best, yellow background = improved over Baseline.

(a) RAFT-Stereo

Method	KITTI-12		KITTI-15		Middlebury		ETH3D	
	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)
Baseline	0.917	95.3	1.196	94.5	**1.075**	**94.6**	0.300	98.8
GeomAug	0.884	**95.6**	**1.088**	**94.7**	1.165	93.1	0.291	99.0
AuxTask	0.934	95.3	1.152	94.2	1.128	94.1	0.285	99.0
UpdReg	**0.879**	95.5	1.109	94.6	1.134	93.7	0.280	98.9
Joint	0.915	95.3	1.109	94.5	1.201	93.4	**0.279**	**99.1**

(b) IGEV-Stereo

Method	KITTI-12		KITTI-15		Middlebury		ETH3D	
	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)
Baseline	1.052	94.2	1.204	93.8	1.054	94.2	0.360	98.3
GeomAug	1.029	94.3	1.142	94.3	**0.909**	**95.6**	**0.297**	**98.7**
AuxTask	1.041	94.4	1.217	93.9	0.924	94.9	0.304	98.6
UpdReg	1.043	94.1	1.187	94.0	0.961	95.2	0.334	98.6
Joint	**0.998**	**94.7**	**1.133**	**94.4**	1.053	94.9	0.329	98.6

(c) Selective-IGEV

Method	KITTI-12		KITTI-15		Middlebury		ETH3D	
	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)
Baseline	1.076	93.9	1.299	93.3	0.990	94.7	**0.333**	98.4
GeomAug	1.037	94.1	1.217	93.8	0.919	95.2	0.345	**98.5**
AuxTask	1.040	94.1	1.271	93.4	0.928	95.2	0.336	**98.5**
UpdReg	0.959	**94.8**	**1.184**	94.3	0.962	95.4	0.336	98.4
Joint	**0.920**	**95.3**	1.246	**94.6**	**0.861**	**95.8**	0.377	**98.5**

(d) DLNR

Method	KITTI-12		KITTI-15		Middlebury		ETH3D	
	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)	EPE (px)	$< 3\,\text{px}$ (%)
Baseline	1.316	93.5	1.747	93.0	1.181	94.5	0.803	97.5
GeomAug	1.189	94.5	1.415	93.6	1.142	95.2	0.394	98.7
AuxTask	1.256	93.7	1.456	92.9	**1.136**	**94.9**	0.421	98.7
UpdReg	**1.107**	**94.8**	**1.238**	**93.9**	1.282	93.8	**0.358**	**98.8**
Joint	1.183	94.6	1.298	93.8	1.226	94.5	0.410	98.4

racy across the image, and $<$ 3 px (%), which reports the percentage of pixels whose disparity error is below 3 pixels.

As indicated in Table 1, we vary the geometry-oriented data augmentation (*GeomAug*) usage probability ρ_{aug} and the Blob-mode ratio γ_{blob}, with the Ribbon-mode ratio being $1 - \gamma_{\text{blob}}$. The best overall performance is achieved with $\rho_{\text{aug}} = 0.1$ and $\gamma_{\text{blob}} = 0.5$, which yields the lowest EPE on KITTI-12 (0.884px), KITTI-15 (1.088px) and ETH3D (0.291px), and consistently high accuracy ($<$ 3 px). We use this configuration in our following experiments as it maintains a balanced variety between Blob and Ribbon perturbation with an efficient strength.

Table 2 reports results for varying loss weight configurations $(\lambda_{\text{obj}}, \lambda_{\text{edge}})$ in auxiliary multi-task context supervision (*AuxTask*). We observe that $(0.2, 0.2)$ achieves the best performance on Middlebury and ETH3D (1.128px and 0.285px) while $(0.2, 0)$ excels on both KITTI datasets (0.882px and 1.099px). As the configuration $(0.2, 0.2)$ offers the optimal trade-off across datasets, with a relatively acceptable performance on KITTI datasets, we take this moderate weight configuration of both tasks to reinforce context understanding without overly diminishing the reliance on geometric cues.

For the depth-update regularization (*UpdReg*) loss weighted by λ_{var} (Table 3), $\lambda_{\text{var}} = 0.1$ achieves the best EPE on both KITTI datasets (0.879px and 1.109px) and ETH3D (0.280px). We choose this small regularization weight (0.1) in follow-up experiments as it provides stability without excessively constraining depth updates.

5.3. Quantitative Results and Analysis

Table 4 presents the quantitative results for four baseline models trained with *GeomAug*, *AuxTask*, and *UpdReg*, as well as their combination (*Joint*), across KITTI-2012, KITTI-2015, Middlebury, and ETH3D. Figure 8 shows the disparity maps for RAFT-Stereo and IGEV-Stereo comparing the baseline with the four training strategies. *Baseline* refers to vanilla training with the augmentation scheme reported in each model's original literature.

Effect of GeomAug. Across all models and datasets, *GeomAug* tends to yield improvements, with particularly pronounced gains for IGEV-Stereo. This aligns with IGEV-Stereo's design, which emphasizes constructing a powerful cost volume through CGEV and benefits further when geometric cues are enriched during training, enabling more robust spatial correspondences. On Middlebury and ETH3D, where geometric complexity and fine details are emphasized, the *GeomAug*-trained IGEV-Stereo achieves the best results (EPE reduced from 1.054px to 0.909px and from 0.360px to 0.297px, respectively). On KITTI datasets, it ranks second-best while still showing consistent gains, highlighting stronger generalization under both high-precision and large-baseline regimes.

Effect of AuxTask. *AuxTask* delivers the most notable gains on Selective-IGEV and DLNR. Both architectures devote significant capacity to preserving high-frequency details and local geometry, which can sometimes bias the network toward fine structure at the expense of leveraging global semantic priors. The auxiliary tasks in our training rebalance this trade-off by reintroducing global context into disparity refinement. This results in improved boundary sharpness and reduced mismatches in textureless areas, effects most evident on indoor datasets Middlebury (EPE of Selective-IGEV improves from 0.990px to 0.928px) and ETH3D (EPE of DLNR improves from 0.803px to 0.421px), where semantic priors help disambiguate repetitive or low-texture regions.

Effect of UpdReg. *UpdReg* consistently improves the performance in KITTI-12 and KITTI-15 across all four models. KITTI's large-scale outdoor driving scenes exhibit a significant domain gap relative to Scene Flow's synthetic content, especially in disparity range and spatial scale. Our update regularization acts to stabilize the iterative refinement process, preventing overfitting to Scene Flow's smaller-scale statistics and allowing more controlled updates when confronted with large, real-world disparities.

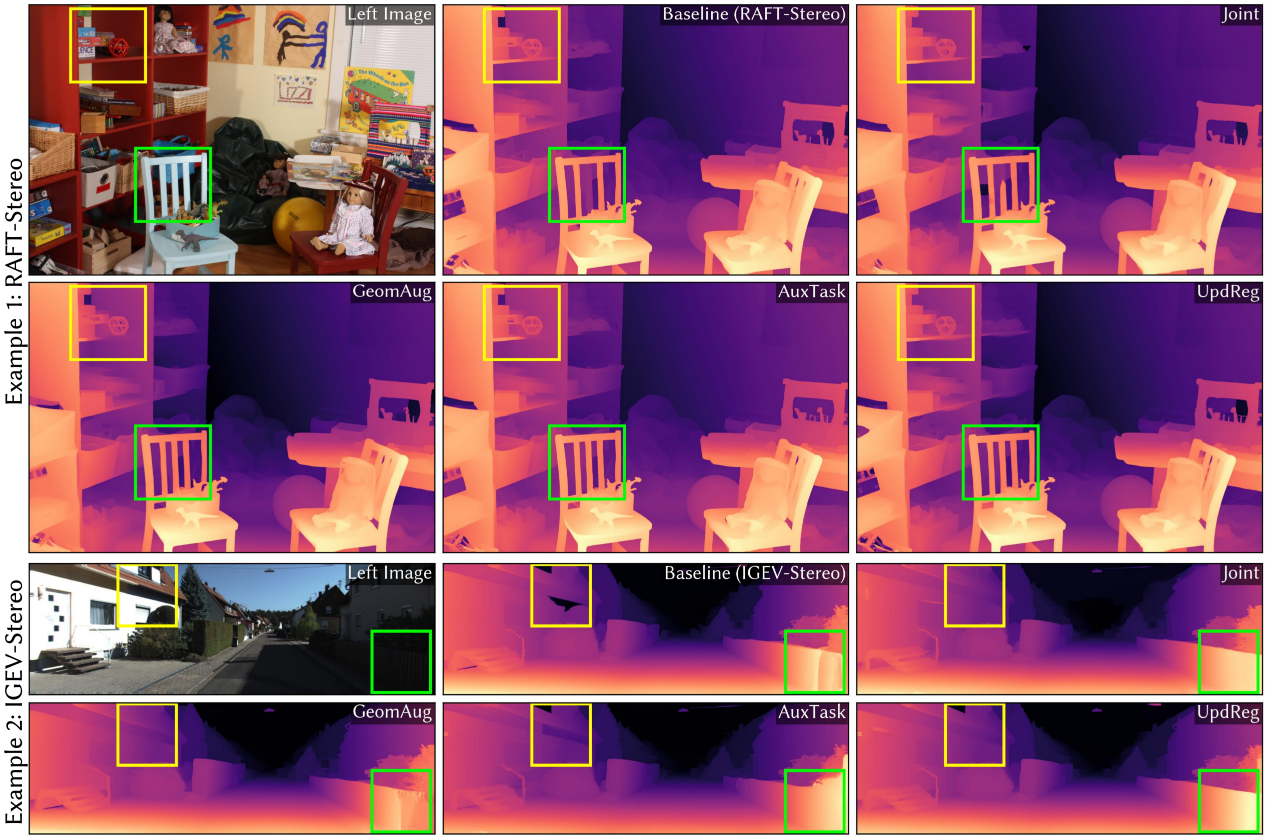

Figure 8. Qualitative results on Middlebury and KITTI 2012 for RAFT-Stereo and IGEV-Stereo with the baseline and our four training strategies(GeomAug, AuxTask, UpdReg and their combination, Joint). Each panel shows the predicted disparity map.

This stability benefit is particularly valuable in architectures with strong iterative backbones such as DLNR using a Multiscale Decouple LSTM, where overly aggressive updates can otherwise degrade cross-domain performance. With *UpdReg*, we obtain the best results: EPE drops from 1.316px to 1.107px on KITTI-12, from 1.747px to 1.238px on KITTI-15, and from 0.803px to 0.358px on ETH3D.

Mixed results of joint training. While *joint training* can yield optimal results in our implementations, most notably for Selective-IGEV on KITTI-12 and Middlebury (with EPE of 0.920px and 0.861px), it does not consistently improve performance across all datasets and models. We attribute this to a potential tension between the geometry-focused bias of *GeomAug* and the semantic emphasis of *AuxTask*, which may pull feature learning in competing directions. Additionally, compounding strong geometric augmentations with additional semantic supervision can increase the effective noise in training signals. While moderate augmentation promotes robustness, excessive perturbation may hinder the network's ability to settle into a well-generalizing solution, particularly for architectures already sensitive to data distribution shifts. This suggest that *joint training* require architecture-specific tuning of loss weights and augmentation strength to avoid diminishing returns.

6. Conclusion

In this paper, we investigate three lightweight, plug-and-play training strategies to improve generalization in stereo matching by aligning supervision with the functional roles of recurrent modules. Without modifying the pipeline or incurring inference overhead, these strategies consistently reduce the synthetic-to-real gap across benchmarks. Our results indicate that under data scarcity and domain shift, task-aligned training signals can enhance robustness without resorting to complex architectures.

Limitations. Our study is limited to simple strategy instantiations and mainstream RNN-based stereo pipelines, and does not yet cover the full design space. We omit real-capture demonstrations and stress tests under adverse environmental conditions. Future work will expand the design space, explore diverse architectures, and target more challenging scenarios.

Acknowledgments

This research is supported by the National Science Foundation of China (62322217), Innovation & Technology Fund (ITP/062/24AP, MHP/313/24) and Research Grants Council of Hong Kong (GRF 17208023). We thank Dr. Tengfei Jiang from SHINING 3D Tech Co., Ltd. for discussions.

References

[1] Filippo Aleotti, Fabio Tosi, Pierluigi Zama Ramirez, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Neural disparity refinement for arbitrary resolution stereo. In *2021 International Conference on 3D Vision (3DV)*, pages 207–217. IEEE, 2021. 2

[2] Luca Bartolomei, Fabio Tosi, Matteo Poggi, and Stefano Mattoccia. Stereo anywhere: Robust zero-shot deep stereo matching even where either stereo or mono fail. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1013–1027, 2025. 3

[3] Changjiang Cai, Matteo Poggi, Stefano Mattoccia, and Philippos Mordohai. Matching-space stereo networks for cross-domain generalization. In *2020 International conference on 3D vision (3DV)*, pages 364–373. IEEE, 2020. 1

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2

[5] Tianyu Chang, Xun Yang, Tianzhu Zhang, and Meng Wang. Domain generalized stereo matching via hierarchical visual transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9559–9568, 2023. 2

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 5

[7] 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 arXiv:2501.08643*, 2025. 1, 3

[8] Jean Gallier. *Geometric methods and applications: for computer science and engineering*. Springer Science & Business Media, 2011. 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 6

[10] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1

[11] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3273–3282, 2019. 1

[12] 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 arXiv:2411.14053*, 2024. 3

[13] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. Defom-stereo: Depth foundation model based stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21857–21867, 2025. 3

[14] Liting Jiang, Feng Wang, Wenyi Zhang, Peifeng Li, Hongjian You, and Yuming Xiang. Rethinking the key factors for the generalization of remote sensing stereo matching networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 1

[15] 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 the IEEE international conference on computer vision*, pages 66–75, 2017. 3

[16] Hyungtae Lee, Yan Zhang, Heesung Kwon, and Shuvra S Bhattacharyya. Exploring the potential of synthetic data to replace real data. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1005–1011. IEEE, 2024. 1

[17] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2

[18] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16263–16272, 2022. 2, 3

[19] Kunhong Li, Longguang Wang, Ye Zhang, Kaiwen Xue, Shunbo Zhou, and Yulan Guo. Los: Local structure-guided stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19746–19756, 2024. 2

[20] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2, 3, 6

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 6

[22] 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 the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 1, 3, 6

[23] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 6

[24] Yikun Miao, Meiqing Wu, Siew Kei Lam, Changsheng Li, and Thambipillai Srikanthan. Hierarchical object-aware dual-level contrastive learning for domain generalized stereo matching. *Advances in Neural Information Processing Systems*, 37:132050–132076, 2024. 1

[25] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16: 100258, 2022. 3

[26] Feiyang Pan, Pengtao Wang, Lin Wang, and Lihong Li. Multi-view stereo vision patchmatch algorithm based on data augmentation. *Sensors*, 23(5):2729, 2023. 3

[27] Zhibo Rao, Yuchao Dai, Zhelun Shen, and Renjie He. Rethinking training strategy in stereo matching. *IEEE Transactions on Neural Networks and Learning Systems*, 34(10): 7796–7809, 2022. 1

[28] Zhibo Rao, Bangshu Xiong, Mingyi He, Yuchao Dai, Renjie He, Zhelun Shen, and Xing Li. Masked representation learning for domain generalized stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5435–5444, 2023. 2

[29] 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 *German conference on pattern recognition*, pages 31–42. Springer, 2014. 1, 6

[30] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 1, 6

[31] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, 128(4):910–930, 2020. 2

[32] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1, 2, 3, 6

[33] Fabio Tosi, Luca Bartolomei, and Matteo Poggi. A survey on deep stereo matching in the twenties. *International Journal of Computer Vision*, 133(7):4245–4276, 2025. 1

[34] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 2

[35] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 2, 3, 6

[36] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *arXiv preprint arXiv:2501.09898*, 2025. 1, 3

[37] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Song Wang, and Lili Ju. Semantic stereo matching with pyramid cost volumes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[38] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12981–12990, 2022. 1, 2

[39] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 2, 3, 5, 6

[40] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[41] Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. Igev++: iterative multi-range geometry encoding volumes for stereo matching. *arXiv preprint arXiv:2409.00638*, 2024. 2, 3

[42] Hongbin Xu, Zhipeng Zhou, Yu Qiao, Wenxiong Kang, and Qiuxia Wu. Self-supervised multi-view stereo via effective co-segmentation and data-augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3030–3038, 2021. 3

[43] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 4

[44] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 636–651, 2018. 2

[45] Chenghao Zhang, Gaofeng Meng, Bing Su, Shiming Xiang, and Chunhong Pan. Monocular contextual constraint for stereo matching with adaptive weights assignment. *Image and Vision Computing*, 121:104424, 2022. 2

[46] Jiawei Zhang, Xiang Wang, Xiao Bai, Chen Wang, Lei Huang, Yimin Chen, Lin Gu, Jun Zhou, Tatsuya Harada, and Edwin R. Hancock. Revisiting domain generalized stereo matching networks from a feature consistency perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13001–13011, 2022. 2

[47] Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, and Yong Zhao. High-frequency stereo matching network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1327–1336, 2023. 6