LiteMatch: Lightweight Zero-Shot Stereo Matching via Cost Volume Stabilization

Md Raqib Khan[1], Santosh Kumar Vipparthi[2], and Subrahmanyam Murala[1]

[1] CVPR Lab, Trinity College Dublin, The University of Dublin, Dublin, Ireland
[2] CVPR Lab, Indian Institute of Technology Ropar, Rupnagar, Punjab, India
khanmd@tcd.ie

Abstract. Despite rapid progress in learning-based stereo matching, high accuracy is often achieved at the cost of heavy backbones and computationally intensive 3D cost volume processing, resulting in substantial memory and runtime overhead. More critically, these methods frequently struggle to generalize across domains, limiting their practical deployment. We present *LiteMatch*, a lightweight stereo matching framework that achieves strong zero-shot generalization through cost volume stabilization-without expensive 3D convolutions. LiteMatch employs two complementary encoders: a Cross-View Correspondence Encoder (CVCE) to capture global cross-view interactions, and a High-Frequency Encoder (HFE) that enhances fine structural details via FFT-based frequency cues. To stabilize the cost volume, we introduce the *Cost Volume Consistency Loss (CVC-Loss)*, a voxel-wise binary cross-entropy objective applied to softmax-normalized cost distributions. By encouraging sharp and unimodal disparity probabilities, CVC-Loss promotes stable cost distributions and enables rapid convergence. A lightweight refinement module further produces sharp full-resolution disparities with low-iteration updates, avoiding heavy recurrent refinement. With a flexible design ranging from 3.36M to 9.58M parameters, LiteMatch achieves exceptional zero-shot generalization, delivering competitive EPE and D1 performance across Scene Flow, KITTI, Middlebury, ETH3D, and DrivingStereo. Our results establish that lightweight architectures can indeed generalize across domains without sacrificing accuracy. Code

1 Introduction

Stereo depth estimation is a cornerstone of 3D perception, with critical applications in autonomous driving, robotics, and augmented reality (AR) [7, 40]. The task demands pixel-accurate disparity maps, particularly at object boundaries, while simultaneously meeting the strict real-time constraints of embedded systems [23].

The advent of deep learning has fundamentally transformed this field. Early models like PSMNet [5] and GA-Net [41] established a powerful paradigm: building a 4D cost volume followed by regularization using 3D convolutional neural networks (CNNs). While effective at encoding geometry, this approach incurs prohibitive computational and memory costs due to the cubic complexity of

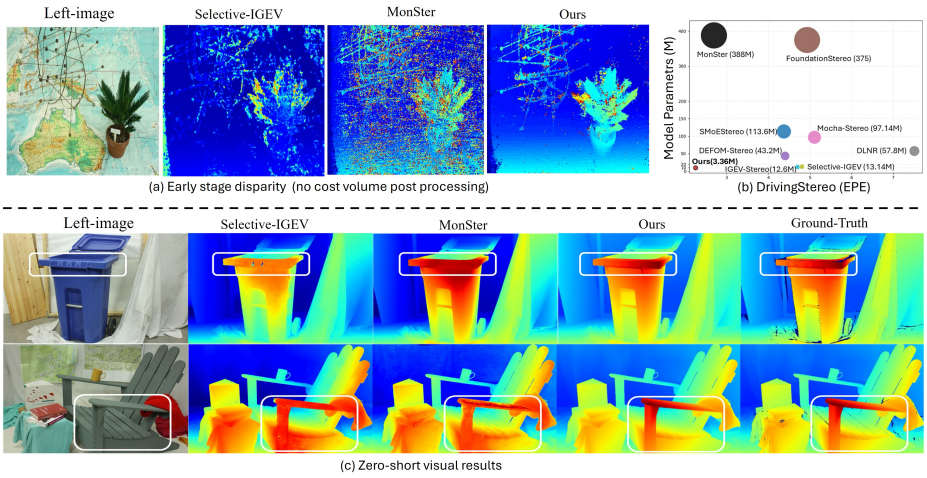

Fig. 1: Zero-shot cross-domain generalization. (a) & (c) LiteMatch produces cleaner disparity maps than state-of-the-art methods on challenging real-world scenes, (b) despite having significantly fewer parameters. This demonstrates superior efficiency and robustness to domain shift.

3D convolutions. This limitation has motivated the exploration of more efficient paradigms. RAFT-Stereo [17] pioneered a different strategy, replacing volumetric processing with iterative refinement driven by a high-resolution correlation pyramid. This achieved scalability but traded one bottleneck for another: its lack of a strong geometric prior led to slow convergence and persistent errors in low-texture regions [35]. Subsequent works like IGEV-Stereo [35] and Selective-IGEV [31] addressed this by hybridizing the approaches, using lightweight 3D cost volumes to generate robust geometry encodings for initializing iterative modules. Nevertheless, this hybrid paradigm still relies on 3D convolutions, preserving significant computational overhead.

Alternatively, a second paradigm leverages massive foundation models and monocular depth priors [8, 12, 34]. While achieving impressive performance, these methods require hundreds of millions of parameters and exhibit slow inference speeds that preclude real-time deployment.

We identify a critical, underlying flaw common to these diverse state-of-the-art approaches: *a fundamental instability in the early feature matching process.* As visualized in Fig. 1 (a), even models with large, pre-trained backbones produce noisy and ambiguous cost volumes. This initial instability creates a dependency cycle: noisy cost volumes necessitate either heavy 3D regularization or dozens of iterative refinement steps to recover accurate disparities. Consequently, the field has predominantly focused on designing increasingly powerful corrective modules (large 3D CNNs, recurrent GRUs), while the root cause of the noise remains unaddressed. This post-processing overhead constitutes the primary bottleneck for efficiency and real-time deployment.

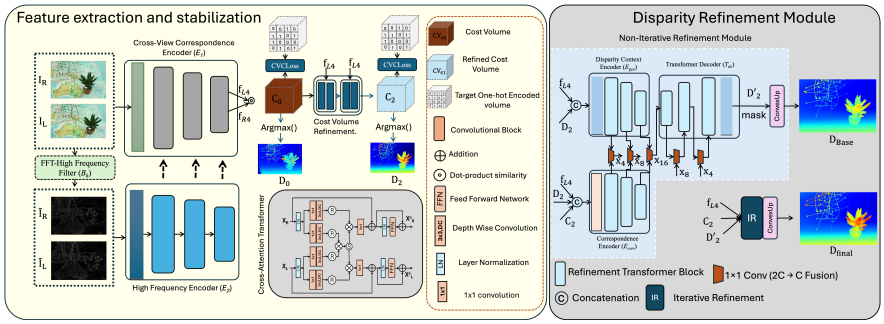

Fig. 2: Overview of LiteMatch. Our framework employs a progressive two-stage scheme: (1) coarse training jointly optimizes the feature extractor and cost volume using our CVC-Loss to establish stable cost volume representations; (2) refinement training freezes Stage 1 components and trains only the transformer-based refinement module with L1 loss to recover fine details. The default model performs efficient single-pass inference, with iterative refinement (IR) extension for comparison.

This analysis leads to a fundamental question: *Can we achieve robust, real-time stereo matching by stabilizing the feature representation and cost volume at the source*, thereby eliminating dependency on 3D convolutions, large pre-trained networks, and external monocular priors?

To address this challenge, we introduce LiteMatch, a novel stereo matching framework designed to prevent cost volume instability at its source. Rather than developing more powerful regularizers, LiteMatch ensures feature stability and cost volume sharpness from the beginning. This principled approach eliminates the need for 3D convolutions or large backbones, establishing an improved balance between performance and complexity. Our main contributions are summarized as follows:

- We propose **LiteMatch**, a lightweight zero-shot stereo matching framework that operates without large backbone networks or monocular depth prior guidance.
- We introduce a frequency-aware stereo representation that integrates prominent correspondence features with high-frequency structural cues, improving disparity estimation in structurally ambiguous regions.
- We propose CVC-Loss, a distribution-level supervision strategy that stabilizes cost volume representations by enforcing inter-bin contrast across disparity hypotheses, producing sharper and more discriminative initial predictions.
- We design a lightweight disparity refinement framework that achieves strong single-pass performance and further attains state-of-the-art accuracy when extended with iterative refinement.

We evaluate LiteMatch extensively on five benchmark datasets (KITTI 2012/2015, Middlebury, ETH3D, and DrivingStereo). Results demonstrate that LiteMatch

achieves superior performance and cross-domain generalization despite its lightweight design, with comprehensive ablation study validating the contribution of each core component of our proposed approach.

2 Related Work

Cost-Volume Regularization and Distribution Modeling. Cost volume construction and regularization remain central to modern stereo matching. Early deep stereo methods relied heavily on stacked 3D CNNs to suppress noisy correlations and enforce spatial consistency, achieving strong accuracy at the expense of computational cost. To improve efficiency, several works have explored alternative cost-volume processing strategies. CFNet [25] leverages multi-resolution cost volumes to better handle large disparity ranges. ACFNet [42] promotes unimodal disparity distributions through adaptive filtering mechanisms. Bi3D [1] reformulates stereo estimation as a sequence of binary disparity threshold classification problems for real-time inference. More recently, ADL [36] models disparity boundary uncertainty using Laplacian mixture distributions to better capture multi-modal ambiguities. These approaches reduce computational overhead or improve uncertainty modeling, yet most methods still depend on either volumetric processing, iterative refinement, or additional mechanisms to mitigate ambiguous cost responses. Designing efficient strategies that maintain stable and well-structured disparity distributions without heavy regularization remains an active research direction.

Feature Extraction Architectures. Feature representation quality strongly influences the reliability of stereo matching. Recent works increasingly adopt large pretrained backbones or monocular priors to enhance cross-domain robustness [3, 8, 34, 38]. While effective, such approaches typically incur substantial computational and memory overhead. Lightweight alternatives aim to improve efficiency by simplifying backbone design or increasing spatial downsampling (e.g., DLNR [43]), which may reduce fine-grained detail preservation and amplify ambiguity in low-texture or repetitive regions. Balancing efficiency and feature fidelity remains a key challenge in lightweight stereo architectures, particularly when striving for robust generalization across diverse domains.

Iterative Refinement Paradigms. Iterative update frameworks have become dominant in high-accuracy stereo systems. RAFT-Stereo [17] introduced recurrent disparity refinement using correlation volumes and gated update operators. Subsequent works enhance this paradigm through normalization techniques (DLNR [43]), motif-based attention mechanisms (MoCha-Stereo [6]), frequency-adaptive fusion strategies (Selective-IGEV [31]), and integration of monocular priors during refinement (MonSter [8]). These approaches achieve state-of-the-art performance but often require multiple refinement iterations, increasing inference latency. An alternative design is An alternative design is HITNet [27], which avoids explicit 3D cost volumes by employing hierarchical tile-based refinement with planar patch propagation and geometric warping. This structured geometric reasoning enables real-time performance with a compact parameter count.

Across both volumetric and geometric paradigms, a fundamental trade-off persists between computational efficiency and cross-domain generalization. Architectures optimized for lightweight inference often simplify feature extraction or refinement, potentially weakening robustness in ambiguous or unseen environments. Conversely, models designed for strong generalization frequently rely on large backbones, iterative updates, or heavy regularization, increasing computational cost. Balancing these competing objectives remains a central challenge in modern stereo matching.

3 Methodology

An overview of our framework is illustrated in Figure 2. We propose LiteMatch, a lightweight and efficient stereo matching framework that achieves high performance without relying on large backbone or depth prior. LiteMatch follows a two-stage design comprising three core components: (1) a dual- branch encoder for stable and geometry-aware feature extraction, (2) the CVC-Loss to stabilize the cost volume during early stages of training, and (3) disparity refinement module is non-iterative for faster inference, but can be optionally extended with an iterative process for fair SOTA comparison. Importantly, our iterative refinement achieves comparable performance with fewer iterations than existing methods.

Pipeline Overview: Given a rectified stereo pair $I_L, I_R \in \mathbb{R}^{H\times W\times 3}$, LiteMatch estimates the final disparity map $D_{\text{final}} \in \mathbb{R}^{H\times W}$ through a two-stage training scheme. In Stage-1, the initial cost volume C_0 is computed using correlation between features f_{L4} and f_{R4} from the dual-encoder, and C_2 is calculated through cost volume refinement (Sec. 3.2). In Stage 2, the Stage-1 parameters are frozen to ensure stability, and a disparity refinement module is trained to take refined C_2 and left feature f_{L4} as inputs, generating a high-resolution disparity map D_{base} and its iterative refinement D_{final}.

3.1 Feature Extraction (Dual Encoders)

Feature quality critically affects stereo matching performance, yet existing backbones face two core limitations. On one hand, convolutional networks [31] provide strong local priors but are limited by small receptive fields, causing oversmoothing in low-texture regions. On the other hand, transformer-based encoders [13, 14, 15, 43] capture global context but lose spatial precision due to global attention. To overcome these limitations, LiteMatch employs a dual encoder that brings together two complementary inductive biases, a Cross-view Correspondence Encoder (E_1) for long-range correspondence modeling, and a High-Frequency Encoder (E_2) for edge and texture preservation [29]. Fusing these representations yields stereo features that are both globally coherent and spatially sharp, improving disparity prediction under diverse conditions.

Cross-view Correspondence Encoder (E_1): It uses a multi-head cross-attention (CA) mechanism to establish dense correspondence between the left

($\mathbf{X}_L$) and right ($\mathbf{X}_R$) features, where $\mathbf{X}_{L/R} \in \mathbb{R}^{B\times C\times H\times W}$. We generate query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) tensors from the input features:

$$\mathbf{Q} = \varphi_3(\psi_1(\mathrm{LN}(\mathbf{X}_L))), \quad \mathbf{K} = \varphi_3(\psi_1(\mathrm{LN}(\mathbf{X}_R))) \tag{1}$$

$$\mathbf{V}_{L/R} = \varphi_3(\psi_1(\mathrm{LN}(\mathbf{X}_{L/R}))) \tag{2}$$

where ψ_1 and φ_3 are 1×1 and depth-wise 3×3 convolutions, respectively. The left-to-right attention map $\mathcal{A}$ is computed as:

$$\mathcal{A} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\cdot\mathbf{K}^\top}{\sqrt{\alpha}}\right), \tag{3}$$

where α is is a learnable scale controlling the magnitude of $\mathbf{Q}\cdot\mathbf{K}^\top$ prior to softmax. We calculate both views using bidirectional attention, leveraging the transpose $\mathcal{A}^\top$ for right-to-left alignment:

$$\mathbf{X}^{\mathrm{out}}_{L/R} = \psi_1(\mathcal{A}\mathbf{V}_{L/R}) + \mathbf{X}_{L/R} \tag{4}$$

Residual connections and FFNs [39] are applied to $\mathbf{X}^{\mathrm{out}}_{L/R}$ to stabilize training and enhance correspondence awareness.

High-Frequency Encoder (E_2) The High-Frequency Encoder (E_2) is a convolutional network designed to preserve high-frequency features (edges and texture). It processes inputs that have been pre-filtered to isolate high-frequency content, $\tilde{I}_L, \tilde{I}_R = \mathcal{H}(I_L, I_R)$, where $\mathcal{H}(\cdot)$ uses a Fourier-based high-pass filter [26]. The filtering operation is defined as:

$$\tilde{I} = \mathcal{F}^{-1}\big(B_h(\mathcal{F}(I))\big), \tag{5}$$

$$B_h(f_{u,v}) = \begin{cases} 0, & |u| < W/4,\ |v| < H/4, \\ f_{u,v}, & \text{otherwise.} \end{cases} \tag{6}$$

Here, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the FFT and its inverse, respectively, and B_h eliminates the low-frequency components in the central region of the frequency domain. The output features of this encoder are denoted $\mathbf{H}_{L/R}$.

Fusion of Complementary Features To maintain spatial sharpness, we fuse the high-frequency convolutional features ($\mathbf{H}$) with the globally coherent transformer outputs ($\mathbf{X}^{\mathrm{out}}$) using a learnable, adaptive gating mechanism. The gate $\mathbf{G}$ determines the optimal contribution of the high-frequency stream:

$$\mathbf{G}_{L/R} = \sigma\big(\psi_g(\mathrm{Cat}(\mathbf{X}^{\mathrm{out}}_{L/R}, \mathbf{H}_{L/R}))\big) \tag{7}$$

where ψ_g is a 1×1 convolution and σ is the sigmoid activation function. The final fused features ($\mathbf{F}$) are calculated as:

$$\mathbf{F}_{L/R} = \mathbf{X}^{\mathrm{out}}_{L/R} + \lambda\mathbf{G}_{L/R}\odot\mathbf{H}_{L/R}, \tag{8}$$

The scaling factor λ is initialized to zero and learned during training. This adaptive gating mechanism effectively balances global coherence and edge fidelity with negligible computational overhead.

3.2 Cost Volume Stabilization

A major bottleneck in stereo matching is heavy 3D cost volume regularization. LiteMatch eliminates this via two components: a Cost Volume Consistency Loss (CVC-Loss) that directly regularizes the disparity distribution, and a lightweight 2D refinement module that filters noisy correspondences.

Cost Volume Refinement: The refinement from C_0 to C_2 is a lightweight 2D feature-guided process defined as:

$$C_{(n)} = C_{(n-1)} + \phi_i\big([C_{(n-1)}, f_{L4}]\big), \quad i \in \{1,\dots,5\}, \quad n \in \{1,2\} \tag{9}$$

where $[\cdot]$ is channel-wise concatenation and ϕ_i represents a 3×3 convolution followed by LeakyReLU.

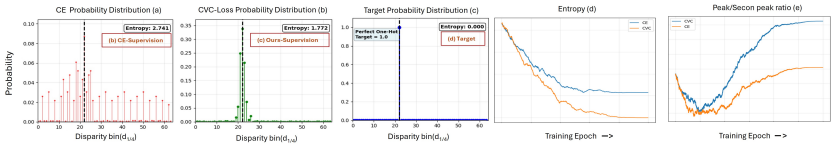

Fig. 3: Cost volume distributions and entropy for a selected pixel. CE (a) shows higher entropy (2.74) than CVC (b, 1.77), with the target one-hot (c) at 0.0. Training dynamics: (d) CVC reduces uncertainty faster, and (e) maintains higher confidence via peak-to-second-peak ratios.

CVC-Loss Formulation: Given cost volume $C \in \mathbb{R}^{H\times W\times D}$ and ground-truth D_{gt}, we obtain probabilities $P_{i,j,d} = \text{Softmax}_d(C_{i,j,d})$. Let $d^*(i,j)$ be the quantized ground-truth bin and Y the corresponding one-hot volume. CVC-Loss applies voxel-wise binary cross-entropy:

$$\mathcal{L}_{\text{CVC}} = -\frac{1}{\sum M \cdot D}\sum_{i,j} M_{i,j} \sum_{d} \Big(Y \log \tilde{P} + (1-Y)\log(1-\tilde{P}) \Big),$$

where $\tilde{P}$ is a stabilized probability and M masks invalid pixels.

For a pixel with ground-truth bin d^*, CVC-Loss decomposes as:

$$\mathcal{L}_{\text{CVC}} = \underbrace{-\log p_{d^*}}_{\text{standard CE}} + \sum_{d\neq d^*} \underbrace{-\log(1-p_d)}_{\text{explicit suppression}} .$$

Thus, CVC-Loss *simultaneously* pushes up the correct probability *and* pulls down every incorrect hypothesis. The gradient for an incorrect bin $k \neq d^*$ is $\frac{\partial}{\partial p_k}(-\log(1-p_k)) = \frac{1}{1-p_k}$, which grows rapidly as $p_k \to 1$, creating an adaptive suppression mechanism. Using $-\log(1-x) \geq x$ for $x \in [0,1)$, we obtain the lower bound:

$$\mathcal{L}_{\text{CVC}} \geq -\log p_{d^*} + (1-p_{d^*}).$$

This shows CVC-Loss *jointly* maximizes the correct probability while minimizing total mass on incorrect bins, producing a provably sharper cost volume than CE alone.

Empirical Behavior and Training Stability: While categorical cross-entropy (CE) is commonly used for disparity classification, it provides weak supervision

for incorrect bins, leading to noisy, multi-modal cost volumes in early training. In contrast, CVC-Loss supplies dense negative supervision across *all* disparity bins via $(1-Y_{i,j,d})\log(1-\tilde{P}_{i,j,d})$, actively suppressing incorrect hypotheses. This yields three measurable benefits (Figs. 3a and 3b): (i) 35.4% lower entropy (1.77 vs. 2.74 for CE); (ii) faster uncertainty reduction during training (Fig. 3d); and (iii) higher confidence via improved peak-to-second-peak ratios (Fig. 3e).

3.3 Disparity Refinement Module

Our refinement strategy emphasizes both high speed and accuracy without requiring mandatory iterative processing. We achieve this through a base non-iterative model, which can be optionally augmented with an IR head for SOTA comparisons

Non-Iterative Disparity Refinement (Base Model): Our base model employs a highly efficient, non-iterative refinement module that enhances disparity in a single forward pass. The input to the module consists of the disparity D_2, left-view image features f_{L4}, and the refined cost volume C_2. The refinement network follows a transformer-based dual-encoder and single-decoder design, leveraging complementary input streams. The Disparity-Context Encoder (E_{geo}) aggregates local image features f_{L4} with the coarse disparity D_2, preserving fine textures and spatial structures around the current prediction. The Correspondence Encoder (E_{corr}), inspired by RAFT's motion encoder [28], fuses the refined cost volume C_2 with D_2 and f_{L4} to capture global correspondence cues within a transformer framework.

The multi-scale outputs from both encoders, $E_{\text{geo},i}$ and $E_{\text{corr},i}$, are fused by a convolutional operator Ψ and passed through a hierarchical transformer decoder T_{de}. This decoder generates an intermediate refined disparity D_2' and a learned convex upsampling mask (M):

$$\text{M}, D_2' = T_{\text{de}}\Psi(\{[E_{\text{geo},i}, E_{\text{corr},i}]\}), \quad i \in \{4, 8, 16\} \tag{10}$$

$$D_{\text{base}} = \mathcal{U}(D_2', \text{M}), \tag{11}$$

where, $\mathcal{U}$ represents the convex upsampling operation. This design generates the full-resolution output D_{base} in a single, efficient pass, enabling real-time disparity estimation.

Iterative Refinement (IR) Extension: For a comprehensive comparison with state-of-the-art iterative methods, we provide an *iterative refinement* (IR) head. This head can be appended to the base model to enable multi-step refinement, following a GRU-based update mechanism similar to that of RAFT-Stereo [17]. The IR head selectively refines uncertain regions using the intermediate disparity, the left feature map, and the cost volume at one-fourth resolution. The refinement process is initialized from the coarse prediction ($\hat{D}_0 = D_2'$) and iteratively updated as:

$$\hat{D}_{t+1} = \text{IR}(\hat{D}_t, f_{L_4}, C_2), \quad t \in \{0, 1, \ldots, n-1\} \tag{12}$$

Table 1: Zero-shot D1-all (%) evaluation on DrivingStereo and in-domain evaluation on Scene Flow. All models are trained exclusively on Scene Flow. MP and BB denote Monocular Prior and Backbone, respectively.

DrivingStereo (Zero-Shot)						Scene Flow (In-Domain)					
Method	Sunny↓	Cloudy↓	Rainy↓	Foggy↓	Avg.↓	**Method**	BB/MP	Pub.	EPE↓	D1 (%)↓	Par(M)↓
Without Monocular Depth Prior						Without Monocular Depth Prior					
CFNet	5.4	5.8	12.0	6.0	7.3	RAFT-Stereo	–	3DV'21	0.53	6.08	11.1
PCWNet	5.6	5.9	11.8	6.2	7.4	DLNR	–	CVPR'23	0.48	5.06	57.4
DLNR	27.1	28.3	34.5	29.0	29.8	IGEV-Stereo	BB	CVPR'23	0.47	5.21	12.6
IGEV-Stereo	5.3	6.3	21.6	8.0	10.3	Selective-IGEV	BB	CVPR'24	0.44	4.98	13.1
Selective-IGEV	7.0	8.0	18.4	12.9	11.1	Mocha-Stereo	BB	CVPR'24	0.41	2.35	97.1
Mocha-Stereo	12.8	27.4	24.6	22.8	21.9	**LiteMatch(Base)**	–	-	0.53	2.29	**3.36**
LiteMatch	**2.01**	**2.05**	**2.44**	**1.51**	**2.00**	**LiteMatch**	–	-	**0.39**	**1.92**	**9.58**
With Monocular Depth Prior						With Monocular Depth Prior					
SMoEStereo	3.3	3.0	6.0	4.7	4.3	DEFOM-Stereo	MP	CVPR'25	0.42	5.57	47.3
DEFOM-Stereo	2.05	2.50	**0.95**	2.68	2.05	SMoEStereo	MP	ICCV'25	0.50	5.55	113
FoundationStereo	3.22	2.81	11.20	2.50	4.93	FoundationStereo	MP	CVPR'25	**0.34**	-	375
MonSter	2.60	2.12	3.08	2.94	2.69	MonSter	MP	CVPR'25	0.37	2.02	388
LiteMatch	**2.01**	**2.05**	2.44	**1.51**	**2.00**	**LiteMatch**	–	-	0.39	**1.92**	**9.58**

Unlike existing SOTA iterative frameworks, our IR converges significantly faster, requiring fewer refinement steps to achieve comparable performance. This dual-mode design provides flexibility to evaluate LiteMatch in both lightweight and high-performance regimes, effectively narrowing the gap with computationally heavier iterative methods.

3.4 Total Loss

Our two-stage training scheme uses distinct objectives for each stage.
Stage 1 applies deep supervision to the cost volumes:

$$L_{\text{stage1}} = \sum_{i=0,2} \left(Smooth_{L1}(D_i - D_{\text{gt4}}) + L_{\text{CVC}}(C_i, D_{\text{gt4}}) \right) \tag{13}$$

where, L_{CVC} denotes the CVC-Loss (see Sec. 3.2), applied to both the initial ($i = 0$) and refined ($i = 2$) cost volumes C_i. We also apply the smooth L_1 loss to the corresponding disparities D_0 and D_2, respectively, where D_{gt4} is the ground truth at one-fourth resolution.
Stage 2 freezes the feature and cost volume modules, and trains the refinement module with a multi-stage regression loss:

$$L_{\text{stage2}} = Smooth_{L1}(D_{\text{base}} - D_{\text{gt}}) + \sum_{j=1}^{n} \gamma^{n-j} \|D_j - D_{\text{gt}}\|_1, \tag{14}$$

Here, D_{base} denotes the base final output, D_j are the intermediate predictions, $\gamma = 0.9$, D_{gt} is the ground-truth disparity, and n indicates the number of iterations. Each loss is used exclusively during its respective training phase. This hybrid supervision ensures stable, geometry-aware learning in Stage 1 and precise, high-resolution disparity in Stage 2, while maintaining real-time performance.

4 Experiments

4.1 Implementation Details

LiteMatch is implemented in PyTorch [21] and follows a two-stage training strategy. Training Pipeline: Our approach follows a two-stage training strategy. In

Table 2: Zero-shot performance on KITTI 2012 and KITTI 2015 (left, **Single-Pass, Iter 1**), and in-domain fine-tuned leaderboard results on ETH3D (right). All models are trained on Scene Flow for zero-shot evaluation.

Zero-Shot (Scene Flow → KITTI)						Fine-Tuned (ETH3D Leaderboard)						
Method	KITTI-15		KITTI-12		FPS	Method	ETH3D-Noc			ETH3D-All		
	EPE	Bad3.0	EPE	Bad3.0			Bad1.0	AvgErr	RMS	Bad1.0	AvgErr	RMS
PSMNet	8.50	16.3	8.0	15.1	3.6	HITNet	2.79	0.20	0.46	3.11	0.22	0.55
CFNet	4.70	-	5.80	-	-	RAFT-Stereo	2.44	0.18	0.36	2.60	0.19	0.42
PCWNet	4.20	-	5.60	-	2.8	IGEV-Stereo	1.12	0.14	0.34	1.51	0.20	0.86
RAFT-Stereo	3.08	13.5	2.74	14.1	11.6	IGEV++	1.14	0.13	0.34	1.58	0.19	0.74
IGEV-Stereo	1.35	6.67	1.16	6.06	16.1	Selective-IGEV	1.23	0.12	0.29	1.56	0.15	0.57
DLNR	15.2	48.4	7.89	30.5	13.0	DEFOM-Stereo	0.70	0.11	0.22	0.78	**0.11**	**0.26**
Selective-IGEV	1.34	6.68	1.23	6.78	14.5	SMoE-Stereo	0.95	0.13	0.29	1.13	0.14	0.36
Mocha-Stereo	1.42	6.80	1.32	6.95	9.13	MonSter	0.44	**0.10**	0.20	0.70	0.13	0.47
LiteMatch (Base)	**1.21**	**5.40**	**1.11**	**5.12**	**22.2**	**LiteMatch**	**0.35**	0.11	**0.19**	**0.52**	0.12	0.31

Table 3: Zero-shot generalization performance on KITTI-2012, KITTI-2015, Middlebury-F, and ETH3D datasets under **default iterative settings**, trained on Scene Flow. Inference time is measured on KITTI-2015 using an RTX A6000 at 384 × 1248 resolution. **Blue bold** indicates the best result and blue indicates the second-best.

Method	Pub.	MP/BB	K12	K15	Middlebury-F		ETH3D		Model Complexity			
			Bad3.0	Bad3.0	EPE	Bad2.0	EPE	Bad1.0	Params	GFLOPs	Mem	Time
Without Monocular Depth Prior												
PSMNet [5]	CVPR-18	-	15.1	16.3	40.51	57.93	-	23.8	5.22	939.6	7.98	332
GWCNet [11]	CVPR-19	-	12.0	11.7	-	32.2	-	14.1	6.43	901.9	9.38	421
RAFT-Stereo [17]	3DV-21	-	5.90	5.86	5.59	19.5	0.36	3.30	11.11	525.9	7.08	353
IGEV-Stereo [35]	CVPR-23	BB	5.19	6.06	4.29	**15.1**	0.33	4.00	12.60	330.6	6.06	410
DLNR [43]	CVPR-23	-	10.6	16.0	6.02	18.3	9.80	23.0	57.38	482.7	6.25	452
Selective-IGEV [31]	CVPR-24	BB	5.64	6.05	5.08	17.6	0.33	6.10	13.14	338.3	6.32	432
Mocha-Stereo [6]	CVPR-24	BB	4.83	6.01	6.08	17.1	0.28	4.02	97.14	739.6	9.79	420
LiteMatch (Base)	-	-	5.12	5.40	4.60	18.1	0.33	5.01	**3.36**	**79.8**	**1.17**	**45**
LiteMatch	-	-	**4.20**	**4.09**	**4.11**	**15.1**	**0.21**	**2.05**	9.58	179.0	1.41	220
With Monocular Depth Prior												
SMoEStereo [32]	ICCV-25	MP	4.22	4.93	4.22	18.9	0.23	2.10	113.6	630	6.70	449
DEFOM-Stereo [12]	CVPR-25	MP	4.29	5.29	6.20	16.6	0.27	2.60	374.4	2236	9.40	458
MonSter [8]	CVPR-25	MP	**3.62**	**4.00**	4.19	15.2	0.24	**2.00**	388.8	2141	7.64	595
LiteMatch	-	-	4.20	4.09	**4.11**	**15.1**	**0.21**	2.05	**9.58**	**179.0**	**1.41**	**220**

Stage 1, we jointly optimize the feature extractor and cost volume modules using our proposed CVC-Loss supervision. Following Selective-IGEV's protocol, we train on Scene Flow for 1 million iterations with batch size 1 using the AdamW optimizer and a OneCycle learning rate schedule (peak 2×10^{-4}), requiring 4-5 days on a single NVIDIA RTX A6000 GPU. In Stage 2, we freeze all Stage 1 components to preserve feature stability and train exclusively the disparity refinement module using regression loss while keeping all hyperparameters consistent. This produces our base model with 3.36M parameters, while the iterative refinement variant (9.58M parameters) additionally trains the IR head during Stage 2.

4.2 Experimental Setup

We evaluate LiteMatch on both synthetic and real-world benchmarks. All models are trained exclusively on the Scene Flow dataset [18], which comprises 35,454 image pairs. We compare LiteMatch against two categories of methods: stereo-only approaches (denoted as "–" or backbone (BB)), including RAFT-Stereo [17],

DLNR [43], IGEV-Stereo [35], Selective-IGEV [31], and Mocha-Stereo [6]; and methods leveraging monocular depth priors (MP), such as DEFOM-Stereo [12], SMoEStereo [32], FoundationStereo [34], and MonSter [8]. The zero-shot generalization performance of LiteMatch is assessed on four diverse datasets: KITTI-2012/2015 [10, 19], Middlebury-F [22], ETH3D [24], and DrivingStereo [37]. We report standard metrics, including End-Point Error (EPE) and D1-all (> 3 px). For KITTI-2012, KITTI-2015, Middlebury-F, and ETH3D, we additionally report Bad X% errors. Inference speed (FPS) is measured at a resolution of 384×1248.

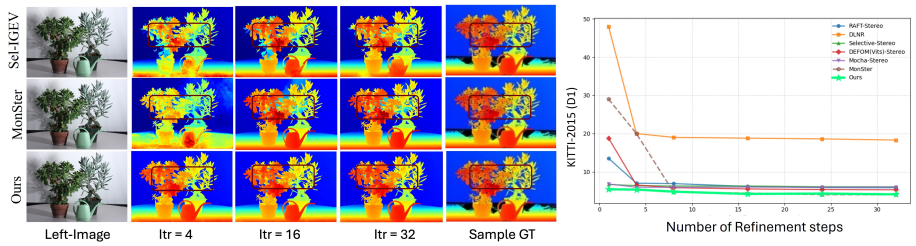

Fig. 4: Iterative Refinement Comparison. Left: Disparity maps at iterations 4, 16, and 32 for MonSter, Selective-IGEV, and LiteMatch, showing convergence and structural clarity. Right: KITTI-2015 D1 error vs. refinement iterations, demonstrating LiteMatch's early stabilization and efficient disparity refinement.

Table 4: Ablation study on model components on Scene flow datasets.

Model	Feature Extraction		CV	Disparity Refinement			CVC-Loss	Params (M)	EPE ↓	D1 (%) ↓
	HFE	CA	Refin.	Geo	Corr	IR				
Baseline	×	×	×	✓	✓	×	×	2.27	0.98	4.48
+ CVC-Loss	×	×	×	✓	✓	×	✓	2.27	0.84	3.30
+ HFE	✓	×	×	✓	✓	×	✓	2.42	0.70	3.18
+ HFE + CA	✓	✓	×	✓	✓	×	✓	2.74	0.65	2.92
+ HFE + Refin. (w/o Corr)	✓	✓	✓	✓	×	×	✓	3.25	0.59	2.80
LiteMatch (Base)	✓	✓	✓	✓	✓	×	✓	**3.36**	**0.53**	**2.29**
LiteMatch	✓	✓	✓	✓	✓	✓	✓	**9.56**	**0.39**	**1.92**

Table 5: Ablation studies trained on Scene Flow. Left: In-domain and cross-domain evaluation (HFE ablation) in terms of D1 (%). Right: Loss function ablation on Scene Flow.

Method	HFE	Scene Flow		DrivingStereo	
		Edge	Non-E	Edge	Non-E
LiteMatch	×	9.45	3.35	10.62	3.78
LiteMatch	✓	**7.21**	**1.42**	**8.24**	**1.85**

Stage 1 Loss	EPE ↓	D1 (%) ↓
Cat. CE + L1	0.73	3.25
Focal (γ=2) + L1	0.64	2.91
CVC-Loss + L1	**0.53**	**2.29**

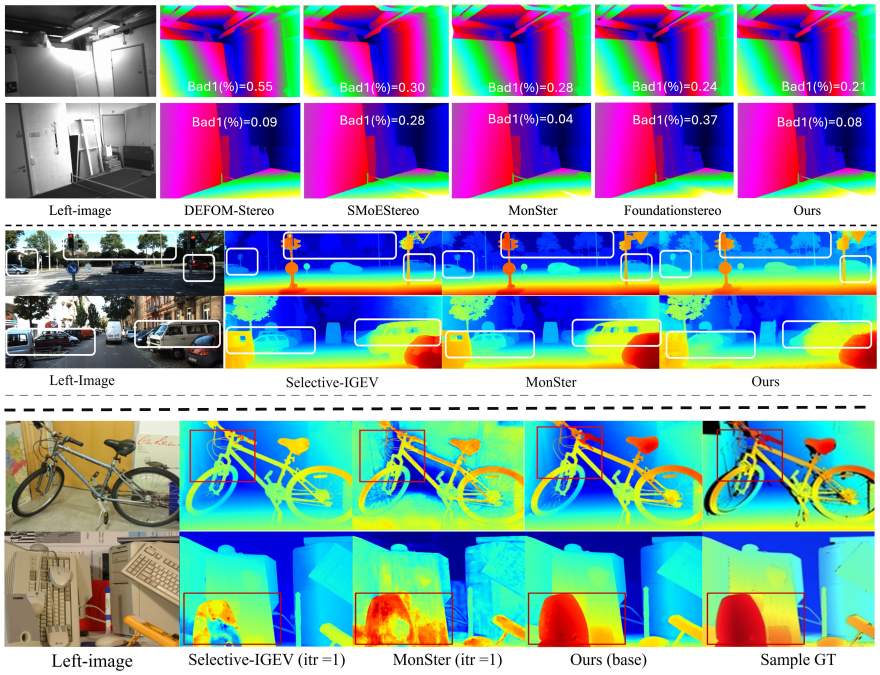

Fig. 5: Qualitative comparison of LiteMatch against state-of-the-art stereo methods, illustrating sharper disparity estimation and cross-domain robustness. Rows 1–2 show results on ETH3D leaderboard samples (fine-tuned), rows 3–4 on KITTI. The final two rows present outputs from the base models and single-iteration refinements of the respective methods for reference.

Comparison with Foundation Models. Recent ETH3D top-performers rely on heavy hybrid designs integrating large pre-trained monocular depth foundations (MonSter 388M, FoundationStereo 375M, DEFOM-Stereo 374M) with 3D convolutions and multi-dataset training (6–8 datasets). In contrast, LiteMatch adopts a monocular-prior-free architecture (3.36M/9.58M) trained only on Scene Flow (35,454 pairs) with single-stage ETH3D fine-tuning. Despite this simplicity, LiteMatch achieves state-of-the-art ETH3D (Bad1.0: 0.42%), surpassing FoundationStereo (0.51%) and MonSter (0.44%), while requiring 40× fewer parameters. This confirms pure stereo matching remains highly competitive with foundation-augmented approaches (see Table S1/S2 in supplementary).

4.3 Performance Evaluation

To validate the effectiveness of LiteMatch, we present comprehensive results on in-domain and cross-domain (zero-shot) generalization benchmarks. We begin with in-domain performance on Scene Flow, followed by leaderboard result on ETH3D, and then evaluate zero-shot generalization across multiple real-world datasets. In-Domain Performance on Scene Flow is shown in Table 1 where LiteMatch (9.58M) is 39× smaller than MonSter (388M) yet achieves the low-

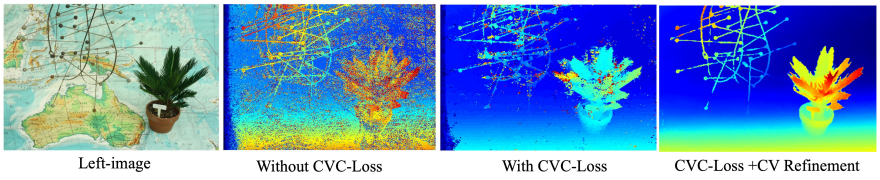

Fig. 6: Visual ablation of CVC-Loss. From left to right: left input image; raw disparity without CVC-Loss (noisy, ambiguous); raw disparity with CVC-Loss (clean, sharp, low-entropy); final stage-1 refined cost volume disparity (clean and preserved details).

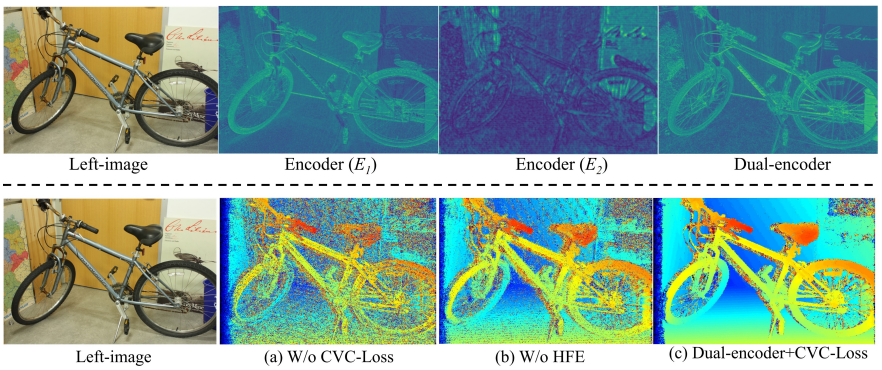

Fig. 7: Top row: Complementary feature activations from the Dual Encoder(E_1)captures smooth, globally coherent regions, while E_2 focuses on sparse, high-frequency edge and texture details; their fusion provides a balanced representation for disparity estimation. Bottom row: Corresponding initial disparity maps raw correlation without CVC-Loss(a) is noisy, without HFE (E_2)(b) is noisy, while fused features and CVC-Loss application produce sharp (c), clean results, reducing the need for heavy post-processing

est EPE (0.39) among all methods without monocular priors and the best D1 (1.92%) overall. It outperforms the large-scale monocular-prior models, demonstrating superior efficiency in stereo matching. Table 2 demonstrates that LiteMatch achieves the best Bad1.0 on both ETH3D-Noc (0.35%) and ETH3D-All (0.52 %), with AvgErr of 0.11/0.12 and RMS of 0.19/0.31. These results position LiteMatch within the top three across all metrics, outperforming MonSter in Bad1.0 while remains competitive on other metrics, demonstrating a strong performance-complexity balance.

We further evaluate LiteMatch's robustness via zero-shot cross-domain trained on Scene Flow generalization, comparing it against both stereo-only and monocular-prior methods. As reported in Table 2, LiteMatch (Base) yields the best EPE (1.21/1.11) and D1 (5.40%/5.12%) among single-iteration stereo-only methods, coupled with the fastest inference speed of 22 FPS on KITTI-2012/2015. These results affirm LiteMatch's *strong generalization to real-world driving scenarios without domain-specific fine-tuning*. Table 3 illustrates the results on various datasets, where LiteMatch sets a new benchmark with the lowest EPE (4.11) and Bad2.0 (15.1%), outperforming MonSter (EPE 4.19, Bad2.0 15.2%) on

Middlebury-F despite being 39× smaller. This performance on high-resolution, dense ground-truth data illustrates LiteMatch's precision in handling complex textures and occlusions. On ETH3D challenging dataset, LiteMatch achieves the best EPE (0.21) and the second-best Bad1.0 (2.05%), closely trailing MonSter (2.00%). The results across varied indoor and outdoor scenes demonstrate LiteMatch's adaptability without relying on external priors. Table 1 presents the results on the DrivingStereo dataset, where LiteMatch achieves the best D1-all (2.00%), averaged over the entire dataset, and sets a new SOTA on the Sunny (2.01%), Cloudy (2.05%), and particularly the Foggy (1.51%) subsets. This dominant performance across diverse weather and illumination conditions *makes it highly suitable for real-world autonomous driving applications.*

Convergence vs. Iterations: We evaluate the impact of iterative refinement on overall performance. Figures 4 illustrate the qualitative (left) and quantitative (right) results, respectively, across different iteration counts (1, 4, . . . , 32). Our LiteMatch demonstrates faster convergence than existing SOTA methods, achieving strong performance even from the first iteration, whereas Selective-Stereo, DLNR, and MonSter begin to converge only after eight or more iterations. Notably, as shown in Figure 5(bottom 2 rows), our base model already outperforms Selective-Stereo and MonSter at a single iteration. *This improvement may be attributed to the proposed stable feature representation and cost volume stabilization modules.*

Performance Vs Complexity Analysis: From Table 3, our model achieves an optimal trade-off between performance and complexity over SOTA methods. Without monocular depth priors, our LiteMatch (Base) model (3.36M params, 79.8 GFLOPs, computed including the FFT filtering module) surpasses all SOTA in inference speed (22.2 FPS vs. 2.8–16.1 FPS for baselines) while maintaining SOTA performance (5.12% Bad3.0 on KITTI-2012, better than IGEV-Stereo's 6.67%). When extended with iterative refinement (LiteMatch), it further closes the performance gap to monocular-prior-based methods (4.09% Bad3.0 vs. MonSter's 4.00%) but remains 2.5× faster (220 ms vs. 595 ms) and 39× lighter (9.58M vs. 388.8M params). This efficiency combined with minimal memory footprint (1.17–1.41 GB vs 7.08–9.4 GB in SOTA) makes our approach *uniquely suited for real-time applications on resource-constrained platforms, from autonomous systems to edge devices.*

4.4 Ablation Study

We perform ablation studies on Scene Flow to evaluate each component of LiteMatch, reporting EPE and D1-all (%). As shown in Table 4, the baseline using only the Cross-view Correspondence Encoder (E1) and non-iterative refinement achieves 0.98 EPE / 4.48% D1, indicating unstable raw correlation. Adding CVC-Loss reduces EPE to 0.84 by stabilizing disparity distributions through dense negative supervision. Incorporating the High-Frequency Encoder (HFE) further improves performance to 0.70 EPE by enhancing structural fidelity at edges and thin regions. Introducing cross-attention (CA) lowers EPE to 0.65, confirming the complementarity of global correspondence modeling and frequency-aware features. With cost volume refinement, EPE decreases to 0.59, while adding the

Correspondence Encoder completes LiteMatch (Base) at 0.53 EPE / 2.29% D1. Enabling the optional iterative refinement (IR) head achieves the best performance of 0.39 EPE / 1.92% D1, demonstrating that stabilized representations support efficient multi-step refinement.

Fig. 6 compares raw and refined disparities with and without CVC-Loss. Without CVC-Loss, the raw disparity is noisy and ambiguous. With CVC-Loss, it becomes clean and sharp, while refinement recovers high-frequency details, confirming that early-stage stabilization reduces the burden on refinement. Figure 7 qualitatively supports these results. E1 captures smooth global structures, whereas HFE emphasizes high-frequency edge details; their fusion yields balanced representations. Without CVC-Loss (Fig. 7(a)), disparity initialization is noisy; without HFE (Fig. 7(b)), boundaries are blurred. The full model (Fig. 7(c)) produces sharp and well-localized disparities, indicating that feature complementarity and cost stabilization reduce ambiguity at initialization.

Table 5 (left) shows that HFE improves edge-region accuracy both in-domain and cross-domain (Scene Flow: 9.45% → 7.21%; DrivingStereo: 10.62% → 8.24%), highlighting improved structural generalization. Table 5 (right) compares Stage 1 supervision strategies: CVC-Loss + L1 outperforms CE and Focal Loss, reducing EPE from 0.64 to 0.53 and D1 from 2.91% to 2.29%. Figure 3 further shows that CVC-Loss produces lower-entropy and more concentrated disparity distributions, confirming its stabilizing effect.

5 Conclusion

This paper presented LiteMatch, a lightweight stereo matching framework that achieves SOTA performance without relying on monocular depth priors or large pre-trained backbones. Our approach addresses the fundamental limitation of cost volume instability through two key contributions: a novel CVC-Loss that stabilizes matching confidence during early training, and a dual-encoder architecture that extracts globally coherent yet spatially precise features. For disparity refinement, we proposed NIR module that refines with single forward pass with an optional IR head serves as a lightweight alternative to SOTA iterative methods. Extensive experiments demonstrate that LiteMatch establishes aan improved balance between performance and complexity. With only 9.58M parameters (39 × fewer than MonSter) our method achieves superior performance on Scene Flow (1.92% D1) and exceptional zero-shot generalization across multiple benchmarks including KITTI, DrivingStereo, Middlebury, and ETH3D. LiteMatch proves that principled architectural design and targeted stabilization strategies can eliminate the dependency on computationally expensive components, enabling high-performance stereo matching for resource-constrained applications.

Acknowledgement:
This research was supported by (d-real) Science Foundation Ireland (Grant 18/CRT/6224). The authors thank the CVPR Lab members at Trinity College Dublin and IIT Ropar for their support.

LiteMatch: Lightweight Zero-Shot Stereo Matching via Cost Volume Stabilization

Supplementary Material

Md Raqib Khan[1], Santosh Kumar Vipparthi[2], and Subrahmanyam Murala[1]

[1] CVPR Lab, Trinity College Dublin, The University of Dublin, Dublin, Ireland
[2] CVPR Lab, Indian Institute of Technology Ropar, Rupnagar, Punjab, India
khanmd@tcd.ie

Overview

This supplementary document provides additional comparison with SOTA, convergence analyses, and additional implementation details for the LiteMatch framework. It includes:

- Section S1: Additional Comparison with SOTA
- Section S2: Convergence Analysis
- Section S3: Additional Implementation Details

S1 Additional Comparison with SOTA

Most recent top-performing methods on ETH3D rely on heavy hybrid designs that integrate large pre-trained monocular depth foundation models (DepthAnythingV2 in DEFOM-Stereo, SMoE-Stereo, MonSter, and FoundationStereo) for feature extraction or initialization, combined with 3D-convolution cost-volume processing and iterative refinement, resulting in parameter counts of 113M–388M (Table S1). These approaches further exploit extensive multi-dataset pre-training and fine-tuning schedules involving 6–8 mixed synthetic/real datasets including(TartanAir [30], CREStereo [16], InStereo2k [2], and ETH3D itself) or even custom 1M-pairs (Table S2), effectively injecting strong monocular priors and domain-specific knowledge that go far beyond pure stereo matching. In contrast, LiteMatch adopts a lightweight, monocular-prior-free architecture (3.36M parameters base, 9.58M with lightweight iterative refinement) and follows a minimal training protocol: pre-training only on the standard Scene Flow dataset (35,454 pairs) followed by single-stage fine-tuning exclusively on ETH3D. Despite this extreme simplicity and the absence of any external monocular or multi-dataset augmentation, LiteMatch achieves state-of-the-art ETH3D, surpassing hybrid foundations such as FoundationStereo (375M parameters, 1M synthetic pairs + DepthAnythingV2 priors), demonstrating that pure, data-efficient stereo matching remains highly competitive with contemporary heavily augmented baselines.

Table S1: Comparison of leading stereo matching architectures based on their core components, highlighting the distinctions between existing methods and ours

Method	Backbone	3D Conv	Iterative Refinement	#param(M)
RAFT-Stereo (3DV'21)	–	No	Yes	11.1
IGEV-Stereo (CVPR'23)	MobileNetV2-100	Yes	Yes	12.6
IGEV++ (TPAMI'2024)	MobileNetV2-100	Yes	Yes	12.4
Selective-IGEV (CVPR'2024)	MobileNetV2-100	Yes	Yes	13.1
DEFoM-Stereo (CVPR'25)	Mono Depth Pre-trained	Yes	Yes	374
SMoE-Stereo (ICCV'2025)	Mono Depth Pre-trained	Yes	Yes	113
MonSter (CVPR'2025)	Mono Depth Pre-trained	Yes	Yes	388
FoundationStereo (CVPR'25)	Mono Depth Pre-trained	Yes	Yes	375
LiteMatch (base)	–	No	No	**3.36**
LiteMatch	–	No	Yes	**9.58**

Table S2: Comparison of training strategies for stereo matching. Our method employs a straightforward pipeline, pre-training on a single synthetic dataset (Scene Flow) and fine-tuning only on the target benchmark (ETH3D). This contrasts with the prevalent trend of using large-scale multi-dataset mixtures for pre-training and/or fine-tuning.

Method	Pre-training	Finetuning Strategy	Finetuning Datasets
RAFT-Stereo [17]	Scene Flow	Single-stage	ETH3D
Selective-IGEV [31]	Scene Flow	Multi-stage	6 datasets
DEFOM-Stereo [12]	Scene Flow	Two-stage	6 datasets
SMoE-Stereo [32]	Scene Flow	Single-stage	4 real datasets
MonSter [8]	8 datasets (mixed)	Multi-stage	Heavy mixture
FoundationStereo [34]	FS80K (∼1M pairs)	Multi-stage	Multiple datasets
Ours	**Scene Flow**	**Single-stage**	**ETH3D only**

S2 Convergence Analysis

Table S3: Bad 3.0 (%) across refinement iterations on KITTI-2015 for models trained on Scene Flow. Our method reaches convergence in fewer iterations.

Iter	RAFT-Stereo	IGEV	DLNR	Selective-IGEV	DEFoM-Stereo	Mocha-Stereo	MonSter	Ours
1	13.50	6.64	48.0	6.68	18.8	6.80	29.0	5.35
4	7.01	6.40	20.0	6.38	6.54	5.95	20.0	4.50
8	6.91	6.28	19.0	6.27	5.84	5.94	4.70	4.30
16	6.17	6.15	18.8	6.27	5.51	5.99	4.40	4.20
24	5.86	6.04	18.6	6.09	5.35	5.95	4.03	4.12
32	5.83	6.03	18.3	6.06	5.29	6.01	4.00	4.09
Converges at.	32	32	32	16	32	8	32	**4**

Table S3 reports the Bad 3.0 (%) across refinement iterations for models trained on Scene Flow and evaluated on KITTI-2015. To quantify convergence speed, we define convergence as the smallest iteration n for which the relative change in EPE between two successive iterations is less than 1%:

$$\frac{|EPE_n - EPE_{n-1}|}{EPE_{n-1}} < 1\%. \tag{1}$$

Under this criterion, our method converges after only four iterations, whereas existing approaches typically require between 8 and 32 iterations. This demonstrates that our refinement process reaches a stable solution significantly faster while maintaining strong performance when generalizing from synthetic (Scene Flow) training to real-world KITTI-2015 datasets.

S3 Additional Implementation Details

Additional training details: For ETH3D fine-tuning, we adapt our training configuration to the benchmark's high-resolution characteristics. Input images are processed at resolution 384×768. We employ a learning rate of 1×10^{-4} upto 40k iterations then reduced to 5×10^{-5} upto 100k , maintaining a batch size of 1, This conservative schedule ensures stable convergence while adapting our SceneFlow-pretrained features.

Additional details for training strategy: We employ two stage training to address optimization challenges in end-to-end stereo learning.

Why the two-stage (Stage-1 → Stage-2) schedule works? Let θ denote Stage-1 parameters (dual-encoder, cost-volume refinement) producing features F_L, F_R, cost volumes C_0, C_2, and coarse disparities D_0, D_2. Let ϕ denote Stage-2 parameters ($\mathrm{E}_{\mathrm{geo}}$, $\mathrm{E}_{\mathrm{corr}}$, decoder, IR). The full joint objective optimized in principle is

$$\min_{\theta,\phi} L(\theta,\phi) \;=\; L_{\mathrm{stage1}}(\theta) \;+\; L_{\mathrm{stage2}}(\theta,\phi), \tag{2}$$

where (following the paper) L_{stage1} contains the Smooth-L_1 terms on D_0, D_2 and the CVC-loss on cost volumes (Sec. 3.2), and L_{stage2} contains the regression losses for the refinement outputs $D_{\mathrm{base}}, D_1, \ldots, D_n$.

Non-stationarity and gradient-variance. The refinement receives input

$$g(x;\theta) = \big(C_2(\theta),\, f_L^4(\theta),\, D_2(\theta)\big).$$

During naive joint training both θ and ϕ are updated simultaneously. Updates to θ change the distribution of $g(x;\theta)$ while ϕ is learning, inducing *non-stationary* training inputs for ϕ [4]. For SGD, the variance of the stochastic gradient for ϕ scales with the variability of g:

$$\mathrm{Var}\big[\nabla_\phi L_{\mathrm{stage2}}(g(x;\theta),\phi)\big] \propto \mathrm{Var}\big[g(x;\theta)\big] \text{ [9]}.$$

High variance slows convergence, increases required damping (smaller LR), and can cause oscillations or collapse of learned refinements [16].

Two-stage decomposition (algorithmic benefits). The practical two-stage schedule:

$$\begin{aligned}
\theta^\star &\approx \arg\min_\theta L_{\mathrm{stage1}}(\theta),\\
\phi^\star &\approx \arg\min_\phi L_{\mathrm{stage2}}(\theta^\star,\phi),\\
&\text{(optional)}\quad (\theta,\phi) \leftarrow \text{few low-LR steps from } (\theta^\star,\phi^\star),
\end{aligned}$$

Our two-stage training strategy yields several important advantages. First, freezing $\theta^\star$ after Stage-1 removes input drift, ensuring that ϕ observes a fixed feature extractor $g(\cdot;\theta^\star)$ throughout refinement. This reduces gradient variance and enables the refinement network to learn consistent residual corrections [20]. Second, solving L_{stage1} first biases θ into an optimization basin characterized by sharp, low-entropy cost volumes a property explicitly enforced by our CVC-loss. Initializing the refinement parameters ϕ within this favorable regime positions the joint parameter set $(\theta^\star, \phi^\star)$ for more effective final finetuning [33]. Third, freezing the backbone prevents destructive gradient updates to the geometric prior encoded by θ, allowing ϕ to learn stable correction operators rather than continually adapting to a moving target [16]. Finally, this staged approach embodies a natural coarse-to-fine curriculum: Stage-1 establishes globally coherent matching structure, while Stage-2 refines local details. This decomposition simplifies each learning task and empirically improves both convergence speed and final accuracy [33].

Analytic sketch (variance reduction). Let $G_\phi(x;\theta,\phi) = \nabla_\phi \ell\big(g(x;\theta),\phi\big)$ be the per-sample gradient. If θ varies across SGD steps, the variance of G_ϕ increases:

$$\mathrm{Var}[G_\phi] = \mathbb{E}_x\big[\|G_\phi\|^2\big] - \|\mathbb{E}_x[G_\phi]\|^2,$$

and the first term grows when $g(x;\theta)$ distribution shifts [4]. Fixing θ eliminates this source of shift, lowering $\mathrm{Var}[G_\phi]$, and thereby improving convergence speed and stability when training ϕ [9].

Training strategy for LiteMatch. Motivated by the success of cascaded architectures, we decompose training into two distinct phases to stabilize learning and specialize module functions. **Stage-1** focuses on learning robust feature correspondence and an initial disparity estimate by optimizing the combined objective L_{stage1} on SceneFlow. **Stage-2** then specializes the refinement network; with the Stage-1 parameters $\theta^\star$ frozen, it is trained solely with L_{stage2} to enhance the resolution and precision of the high-resolution output, preventing the refinement process from interfering with the already-learned robust matching capabilities.

Bibliography

[1] Badki, A., Troccoli, A., Kim, K., Kautz, J., Sen, P., Gallo, O.: Bi3d: Stereo depth estimation via binary classifications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1600–1608 (2020) 4

[2] Bao, W., Wang, W., Xu, Y., Guo, Y., Hong, S., Zhang, X.: Instereo2k: a large real dataset for stereo matching in indoor scenes. Science China Information Sciences **63**(11), 212101 (2020) 16

[3] Bartolomei, L., Tosi, F., Poggi, M., Mattoccia, S.: 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. pp. 1013–1027 (2025) 4

[4] Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM review **60**(2), 223–311 (2018) 18, 19

[5] Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5418 (2018) 1, 10

[6] Chen, Z., Long, W., Yao, H., Zhang, Y., Wang, B., Qin, Y., Wu, J.: Mocha-stereo: Motif channel attention network for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27768–27777 (2024) 4, 10, 11

[7] Chen, Z., Zhao, Y., He, J., Lu, Y., Cui, Z., Li, W., Zhang, Y.: Feature distribution normalization network for multi-view stereo. The Visual Computer **41**(1), 409–421 (2025) 1

[8] Cheng, J., Liu, L., Xu, G., Wang, X., Zhang, Z., Deng, Y., Zang, J., Chen, Y., Cai, Z., Yang, X.: Monster: Marry monodepth to stereo unleashes power. arXiv preprint arXiv:2501.08643 (2025) 2, 4, 10, 11, 17

[9] Faghri, F., Duvenaud, D., Fleet, D.J., Ba, J.: A study of gradient variance in deep learning. arXiv preprint arXiv:2007.04532 (2020) 18, 19

[10] Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012) 11

[11] Guo, X., Yang, K., Yang, W., Wang, X., Li, H.: Group-wise correlation stereo network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3273–3282 (2019) 10

[12] Jiang, H., Lou, Z., Ding, L., Xu, R., Tan, M., Jiang, W., Huang, R.: Defom-stereo: Depth foundation model based stereo matching (2025) 2, 10, 11, 17

[13] Khan, M.R., Kulkarni, A., Phutke, S.S., Murala, S.: Underwater image enhancement with phase transfer and attention. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2023) 5

[14] Khan, M.R., Negi, A., Kulkarni, A., Phutke, S.S., Vipparthi, S.K., Murala, S.: Phaseformer: Phase-based attention mechanism for underwater image

restoration and beyond. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 9618–9629. IEEE (2025) 5

[15] Khan, R., Mishra, P., Mehta, N., Phutke, S.S., Vipparthi, S.K., Nandi, S., Murala, S.: Spectroformer: Multi-domain query cascaded transformer network for underwater image enhancement. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1454–1463 (2024) 5

[16] Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., Liu, J., Fan, H., Liu, S.: Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16263–16272 (2022) 16, 18, 19

[17] Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021) 2, 4, 8, 10, 17

[18] Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: 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. pp. 4040–4048 (2016) 10

[19] Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3061–3070 (2015) 11

[20] Pang, J., Sun, W., Ren, J.S., Yang, C., Yan, Q.: Cascade residual learning: A two-stage convolutional neural network for stereo matching. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 887–895 (2017) 19

[21] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NeurIPS) (2019) 9

[22] Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36. pp. 31–42. Springer (2014) 11

[23] Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International journal of computer vision **47**(1), 7–42 (2002) 1

[24] Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: 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. pp. 3260–3269 (2017) 11

[25] Shen, Z., Dai, Y., Rao, Z.: Cfnet: Cascade and fused cost volume for robust stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13906–13915 (2021) 4

[26] Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain

learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5052–5060 (2024) 6

[27] Tankovich, V., Hane, C., Zhang, Y., Kowdle, A., Fanello, S., Bouaziz, S.: Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14362–14372 (2021) 4

[28] Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020) 8

[29] Wang, J., Du, R., Chang, D., Liang, K., Ma, Z.: Domain generalization via frequency-domain-based feature disentanglement and interaction. In: Proceedings of the 30th ACM international conference on multimedia. pp. 4821–4829 (2022) 5

[30] Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam (2020) 16

[31] Wang, X., Xu, G., Jia, H., Yang, X.: Selective-stereo: Adaptive frequency information selection for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19701–19710 (2024) 2, 4, 5, 10, 11, 17

[32] Wang, Y., Wang, L., Zhang, C., Zhang, Y., Zhang, Z., Ma, A., Fan, C., Lam, T.L., Hu, J.: Learning robust stereo matching in the wild with selective mixture-of-experts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21276–21287 (2025) 10, 11, 17

[33] Weinshall, D., Cohen, G., Amir, D.: Curriculum learning by transfer learning: Theory and experiments with deep networks (2018), https://arxiv.org/abs/1802.03796 19

[34] Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. arXiv preprint arXiv:2501.09898 (2025) 2, 4, 11, 17

[35] Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21919–21928 (2023) 2, 10, 11

[36] Xu, P., Xiang, Z., Qiao, C., Fu, J., Pu, T.: Adaptive multi-modal cross-entropy loss for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5135–5144 (2024) 4

[37] Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 899–908 (2019) 11

[38] Yao, C., Yu, L., Liu, Z., Zeng, J., Wu, Y., Jia, Y.: Diving into the fusion of monocular priors for generalized stereo matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14887–14897 (2025) 4

[39] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022) 6

[40] Zbontar, J., LeCun, Y.: Computing the stereo matching cost with a convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1592–1599 (2015) 1

[41] Zhang, F., Prisacariu, V., Yang, R., Torr, P.H.: Ga-net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 185–194 (2019) 1

[42] Zhang, Y., Chen, Y., Bai, X., Yu, S., Yu, K., Li, Z., Yang, K.: Adaptive unimodal cost volume filtering for deep stereo matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12926–12934 (2020) 4

[43] Zhao, H., Zhou, H., Zhang, Y., Chen, J., Yang, Y., Zhao, Y.: High-frequency stereo matching network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1327–1336 (2023) 4, 5, 10, 11