CURA-Stereo: Cross-Volume Consistency Uncertainty for Fusion-and-Rectification in Iterative Stereo Matching

Qingqing Cheng†
Nanchang Hangkong University
Nanchang, China
2308080400027@stu.nchu.edu.cn

Dongyang Wang†
Nanchang University
Nanchang, China
dongyang0524@email.ncu.edu.cn

Chao He*
Nanchang Hangkong University
Nanchang, China
hechao92918@163.com

Zhen Chen
Nanchang Hangkong University
Nanchang, China
dr_chenzhen@163.com

Xinping Mao
Nanchang Hangkong University
Nanchang, China
mxp01052002@163.com

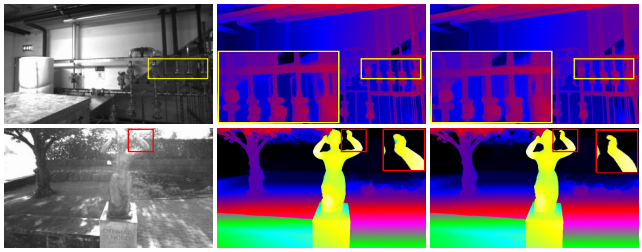
Congxuan Zhang*
Nanchang Hangkong University
Nanchang, China
zcxdsg@163.com

***Abstract*—In recent years, iterative stereo matching has become a widely adopted paradigm for dense disparity estimation by refining predictions with feature warping and recurrent updates. However, in ill-posed regions such as occlusions, weak textures, and depth discontinuities, inaccurate disparity estimates can cause unreliable warping, introducing noisy cues and accumulating errors across iterations. To address this issue, we propose CURA-Stereo, an uncertainty-driven iterative stereo framework that explicitly models alignment reliability inside the refinement loop. Specifically, Cross-Volume Consistency Uncertainty (CVCU) estimates a unified uncertainty map that guides Uncertainty-Conditioned Soft Fusion (UCSF) and Uncertainty-Driven Alignment Rectification (UDAR) to adaptively fuse features and rectify distorted alignments, thereby suppressing error propagation. Extensive experiments demonstrate that CURA-Stereo achieves 0.44 px EPE on Scene Flow, improves KITTI 2012 2-noc to 1.56 and KITTI 2015 D1-fg to 2.36, respectively, and attains 0.32 Bad 2.0 on ETH3D. Without fine-tuning, it further reaches 3.1 on ETH3D and 5.0 on Middlebury at quarter resolution, indicating stable convergence and strong robustness under domain shifts with marginal computational overhead.**

***Index Terms*—Iterative stereo matching, Uncertainty estimation, Feature fusion, Alignment rectification**

I. Introduction

Stereo matching [1]–[3] is a fundamental computer vision [4] problem that estimates dense disparity from rectified stereo pairs, providing reliable depth for 3D reconstruction [5], autonomous driving [6], and robotic navigation [7]. With the advances of end-to-end deep networks, the dominant paradigm has shifted from one-shot cost-volume regression (e.g., PSMNet [1] and GwcNet [8]) to recurrent iterative refinement. Representative methods such as RAFT-Stereo [9] and IGEV-Stereo [10] integrate feature warping with recurrent update operators (e.g., ConvGRU [11]), transforming global correspondence search into iterative local residual updates and achieving strong performance on standard benchmarks.



(a) Left Image (b) IGEV-Stereo (c) CURA-Stereo(Ours)

Fig. 1. Qualitative comparisons on the ETH3D test set. CURA-Stereo yields sharper boundaries and more consistent geometry than IGEV-Stereo in challenging regions.

Despite these advances, iterative stereo matching [12] remains vulnerable in ill-posed regions, such as occlusions, textureless areas, and disparity discontinuities around object boundaries. This is mainly because these frameworks warp right-view features into the left coordinate system using the previous disparity estimate, making the aligned features highly sensitive to disparity errors [3], [9], [10], [13], [14]. In challenging regions, inaccurate disparities can cause misalignment and distortion in the warped features, break cross-view consistency, and introduce noisy cues into the recurrent update unit. Such unreliable signals may be progressively amplified over iterations, leading to error accumulation and degraded convergence stability.

Moreover, existing methods often fail to leverage alignment uncertainty: most uncertainty estimators are computed at a single scale and many fusion strategies are not explicitly reliability-aware, making it difficult to adaptively balance original features and warped features under spatially varying ambiguity [9], [10]. Although prior work [15] leverages feature-consistency cues to alleviate errors in difficult regions, the

† These authors contributed equally to this work.
∗ Corresponding authors: hechao92918@163.com, zcxdsg@163.com.

refinement loop typically lacks uncertainty-driven gating and rectification, which may result in under- or over-correction.

To address these challenges, we propose CURA-Stereo, an uncertainty-driven iterative stereo matching framework that explicitly models alignment uncertainty within the recurrent refinement loop. At each iteration, CVCU estimates a unified uncertainty map U from multi-scale cross-volume inconsistency, by contrasting a global cost volume with a locally aligned cost volume. Guided by U, UCSF predicts pixel-wise fusion weights to softly combine the original and aligned features, while UDAR rectifies the aligned features via uncertainty-modulated corrections to suppress distortion propagation. After each update, CURA-Stereo re-estimates U and uses it to regulate fusion and rectification in the subsequent iteration, thereby mitigating error accumulation in ill-posed regions. Our main contributions are summarized as follows:

- **Multi-scale uncertainty from cross-volume inconsistency.** We estimate alignment uncertainty from multi-scale cross-volume inconsistency and aggregate it into a unified uncertainty map U, improving robustness in ambiguous regions.
- **Uncertainty-conditioned soft fusion for stable refinement.** We propose UCSF, which leverages U to adaptively fuse original and warped features, improving refinement stability under spatially varying uncertainty.
- **Uncertainty-driven alignment rectification.** We propose UDAR to rectify warped features under the guidance of U with lightweight regularization, reducing error accumulation during iterative updates.

II. Related Work

A. Iterative Optimization-Based Stereo Matching

Stereo matching has recently evolved from one-shot cost-volume regression to recurrent iterative refinement. Representative methods such as RAFT-Stereo [9] refine disparities through multi-step residual updates by repeatedly aggregating correlation features with contextual cues [1], [2], [16]–[18]. Building on this paradigm, IGEV-Stereo [10] incorporates geometric priors into an indexable geometry-encoding volume to enhance structural consistency, while PCW-Net [8] improves the initial representation and exploits warping cues to narrow the residual search space, achieving a favorable trade-off between accuracy and generalization.

A key component of these frameworks is to warp right-view features into the left view using disparities estimated in the previous iteration. However, in ill-posed regions such as occlusions, textureless areas, and depth discontinuities, disparity errors can distort the warped features and propagate through recurrent updates, leading to error accumulation over iterations. This motivates reliability-aware regulation of alignment cues to improve robustness and convergence stability.

B. Uncertainty and Confidence Modeling in Stereo Matching

Uncertainty (or confidence) modeling aims to estimate the reliability of stereo predictions and has been widely studied for quality assessment and risk-aware learning [13], [14]. Early works typically [18]–[20] rely on post-hoc confidence measures, while recent approaches perform end-to-end uncertainty estimation via risk-based objectives (e.g., Bayes risk) or uncertainty decomposition. Content-adaptive [2], [21] designs have also been explored to reduce matching ambiguity; for instance, Selective-Stereo [3] modulates iterative updates across regions using selectable operators and attention.

For iterative stereo matching, uncertainty should reflect not only the reliability of disparity predictions but also the iteration-dependent trustworthiness of warping-aligned features. However, most existing cues rely on single-scale or single-source evidence, limiting their ability to capture ambiguities in both fine textures and global structures and to enable explicit uncertainty-conditioned regulation within the recurrent loop. Therefore, a multi-scale reliability signal tightly coupled with the alignment process is desirable. In this work, we derive such uncertainty from cross-volume inconsistencies and use it to guide uncertainty-conditioned fusion and rectification.

C. Feature Warping and Consistency Compensation

Feature warping aligns right-view features to the left-view coordinate system using the current disparity estimate, which enables iterative stereo frameworks to refine disparities via local residual updates [19], [22]. However, warping is highly sensitive to disparity errors: in ill-posed regions [23](e.g., occlusions, weak textures, specular surfaces, and disparity discontinuities), inaccurate estimates can induce misalignment and distortion in the warped features, corrupting the matching representation and propagating errors through recurrent updates [9].

To mitigate warping-induced artifacts, prior work has explored consistency-based compensation. For example, WCG-Net [15] derives attention cues from discrepancies between left-view and warped features and injects a warping correlation volume into the update unit to improve refinement in challenging regions. While effective, such methods often rely on implicit reliability cues and do not explicitly quantify alignment trustworthiness, making compensation difficult to interpret and control. This motivates an explicit reliability signal to regulate the use of aligned features and trigger rectification when needed, thereby suppressing error propagation and stabilizing iterative convergence.

III. Method

As shown in Fig. 2, the proposed CURA-Stereo is built upon IGEV-Stereo as the baseline and follows a three-stage pipeline of "feature extraction–cost volume construction–iterative disparity refinement [3], [9], [10]." Our improvements are introduced in the iterative refinement stage: at each iteration, CVCU estimates U, which guides UCSF fusion and UDAR rectification before ConvGRU residual updates.

A. Multi-Scale Feature & Context Extraction

Given a rectified stereo pair, we employ a weight-sharing MobileNetV2 [24] pretrained on ImageNet to extract a multi-scale feature pyramid $\{F_L^s, F_R^s\}$, where $s \in \{1/4, 1/8, 1/16\}$.

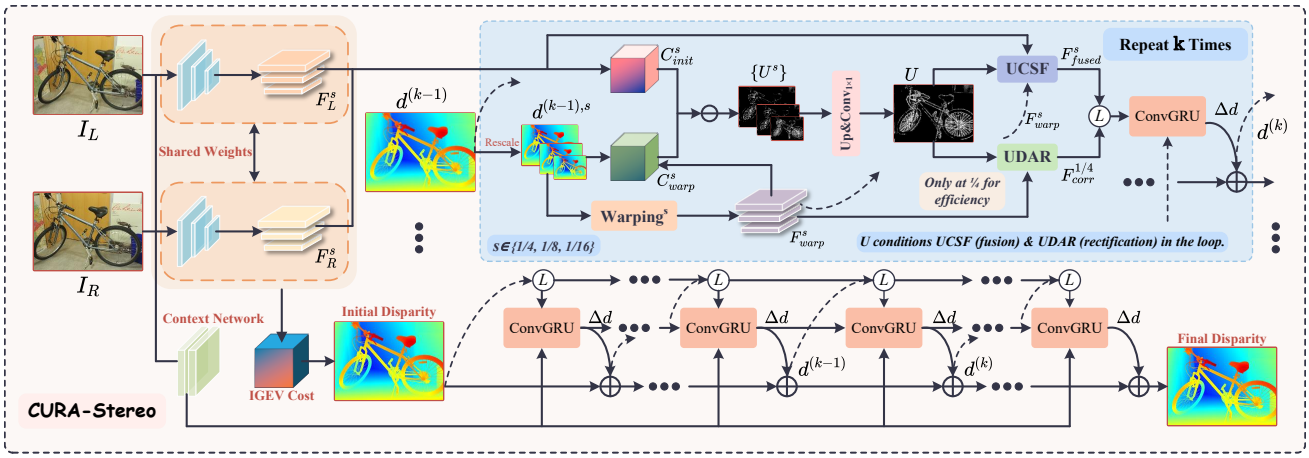


Fig. 2. **Overview of our proposed CURA-Stereo.** CVCU derives a unified uncertainty map U from multi-scale cross-volume inconsistency, which guides UCSF for uncertainty-conditioned feature fusion and UDAR for alignment rectification before recurrent refinement. The uncertainty map is iteratively updated to form a closed-loop control signal that mitigates warping-induced error accumulation.

The extracted features are used to construct the baseline cost volume and regress the initial disparity $d^{(0)}$. They are also fed into CVCU/UCSF for uncertainty estimation and feature fusion, and into UDAR for aligned-feature rectification. In addition, following IGEV-Stereo, the ConvGRU-based update operator takes contextual features from the context branch to support iterative state updates.

B. Scale-Consistent Warping

At the k-th iteration, we rescale the previous disparity $d^{(k-1)}$ according to the scale s to obtain $d^{(k-1),s}$, which is then used to align the right-view features to the left-view coordinate system:

$$F^s_{warp} = \mathrm{Warp}(F^s_R, d^{(k-1),s}). \tag{1}$$

Here, $\mathrm{Warp}(\cdot)$ is implemented via bilinear sampling to achieve sub-pixel alignment. This operation ensures that the geometric displacement on features remains consistent with the disparity magnitude across scales, thereby avoiding systematic bias caused by scale mismatch.

C. Cross-Volume Consistency–Based Uncertainty (CVCU)

Let $\mathrm{Corr}(\cdot)$ denote the operator that constructs a correlation/cost volume along the disparity dimension, consistent with the correlation layer in IGEV-Stereo. At each scale s, we construct two comparable cost volumes. The global cost volume, independent of the current-iteration disparity, is defined as

$$C^s_{init} = \mathrm{Corr}(F^s_L, F^s_R; \mathcal{D}_s), \tag{2}$$

where $\mathcal{D}_s$ denotes the set of global disparity candidates at scale s with cardinality D_s. The locally aligned cost volume is computed by centering at $d^{(k-1),s}$ and correlating within $\Delta d \in [-r, r]$. It is then embedded into the global disparity space by filling the corresponding indices and setting the remaining entries to $-\infty$, yielding C^s_{warp} with the same disparity dimension D_s as C^s_{init}, so that non-local disparities receive negligible probabilities after softmax.

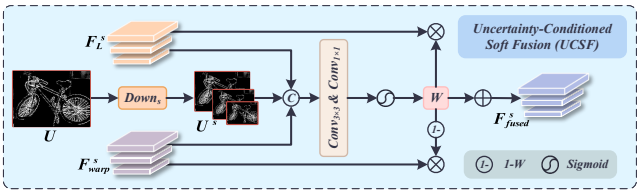


Fig. 3. **Uncertainty-Conditioned Soft Fusion (UCSF).** U^s modulates pixel-wise fusion of F^s_L and F^s_{warp} into F^s_{fused}.

We apply a softmax along the disparity dimension to obtain probabilistic matching distributions:

$$\begin{aligned} \tilde{C}^s_{init} &= \mathrm{Softmax}_d(C^s_{init}), \\ \tilde{C}^s_{warp} &= \mathrm{Softmax}_d(C^s_{warp}). \end{aligned} \tag{3}$$

We quantify the cross-volume inconsistency by computing the discrepancy tensor Δ^s and measuring its variance along the disparity dimension, yielding the scale-specific uncertainty U^s.

$$\Delta^s = \tilde{C}^s_{init} - \tilde{C}^s_{warp}, \tag{4}$$

$$U^s = \mathrm{Var}_d(\Delta^s). \tag{5}$$

Finally, we normalize U^s into a bounded range $\hat{U}^s \in [0, 1]$ via per-image min–max normalization:

$$\hat{U}^s = \frac{U^s - \min(U^s)}{\max(U^s) - \min(U^s) + \epsilon}, \tag{6}$$

where ϵ is added to the denominator for numerical stability.

To obtain a unified control signal that captures ambiguities from fine-grained textures to global structures, we upsample each $\hat{U}^s$ to the $1/4$ resolution and fuse them to produce a unified uncertainty map U:

$$U = \Psi(\{\mathrm{Up}_{1/4}(\hat{U}^s)\}_s). \tag{7}$$

Here, $\Psi(\cdot)$ is implemented as a 1×1 convolution to adaptively aggregate uncertainty cues across scales. The unified uncertainty map U is used to guide both feature fusion (UCSF) and alignment rectification (UDAR) in the refinement loop.

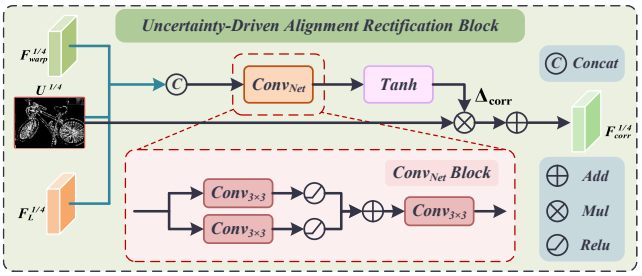


Fig. 4. **Uncertainty-Driven Alignment Rectification (UDAR).** $U^{1/4}$ modulates a lightweight correction to rectify $F_{\text{warp}}^{1/4}$ into $F_{\text{corr}}^{1/4}$.

D. Uncertainty-Conditioned Soft Fusion (UCSF)

Given the unified uncertainty map U at the $1/4$ resolution, we downsample it to each scale as $U^s = \text{Down}_s(U)$ and use it as a conditioning signal to predict pixel-wise fusion weights via a lightweight convolutional network, as shown in Fig. 3:

$$A^s = \text{Conv}_{3\times3}\big(\text{Concat}\big[U^s, F_L^s, F_{warp}^s\big]\big)\,. \tag{8}$$

The fusion weight is obtained by a 1×1 convolution with a Sigmoid activation:

$$W_{uncertain}^s = \sigma(\text{Conv}_{1\times1}(A^s))\,, \tag{9}$$

where $W_{uncertain}^s \in [0,1]$ is a single-channel, pixel-wise map. We compute the fused feature as

$$F_{fused}^s = (1-W_{uncertain}^s)\odot F_{warp}^s + W_{uncertain}^s \odot F_L^s, \tag{10}$$

where $\odot$ denotes element-wise multiplication. This fusion favors F_{warp}^s in low-uncertainty regions while relying more on F_L^s in ambiguous areas, providing more stable features for iterative refinement.

E. Uncertainty-Driven Alignment Rectification (UDAR)

In ill-posed regions, the geometrically aligned feature F_{warp}^s is prone to misalignment and distortion. To alleviate this issue, we propose UDAR, which learns a pixel-wise correction term $\Delta_{correction}$ to rectify F_{warp}^s and adaptively scales the rectification with the uncertainty map U, as shown in Fig. 4.

Since the unified uncertainty map U is fused at the $1/4$ resolution, we set $U^{1/4} = U$. At this scale, UDAR takes $F_L^{1/4}$, $F_{warp}^{1/4}$, and $U^{1/4}$ as inputs and predicts the correction offsets using a lightweight convolutional network:

$$\begin{aligned} \textit{Temp} &= \text{Concat}\Big[F_L^{1/4}, F_{warp}^{1/4}, U^{1/4}\Big], \\ \Delta_{correction} &= \tanh\Big(\text{Conv}_{Net}(\textit{Temp})\Big). \end{aligned} \tag{11}$$

Here, $\Delta_{correction}$ has the same spatial resolution and channel dimension as $F_{warp}^{1/4}$, and the $\tanh(\cdot)$ activation constrains its magnitude to $[-1,1]$ for stable optimization. Conv_{Net} is lightweight and shared across iterations, consisting of two parallel 3×3 Conv+ReLU branches whose outputs are concatenated and fed into a 3×3 output layer.

After obtaining $\Delta_{correction}$, we modulate the correction strength using $U^{1/4}$:

$$F_{corr}^{1/4} = F_{warp}^{1/4} + U^{1/4} \odot \Delta_{correction}. \tag{12}$$

This formulation ensures that the correction magnitude approaches zero in low-uncertainty regions, while allowing larger corrections in high-uncertainty regions.

Loss Function. At the $1/4$ scale, we define an uncertainty-weighted correction loss:

$$L_{correction} = \Big\| U^{1/4} \odot \Big(F_L^{1/4} - F_{corr}^{1/4}\Big)\Big\|_2^2, \tag{13}$$

and add an offset-magnitude regularization term:

$$L_{reg} = \|\Delta_{correction}\|_2^2\,. \tag{14}$$

The overall loss is then given by

$$L = L_{disp} + \lambda_c L_{correction} + \lambda_r L_{reg}, \tag{15}$$

where L_{disp} denotes the primary disparity regression loss, and λ_c and λ_r are weighting coefficients.

F. Iterative Update and Optimization

At iteration k, the ConvGRU-based update operator takes the fused/rectified features together with contextual information to predict a disparity residual and update the estimate to $d^{(k)}$. For residual estimation, the UDAR-rectified feature $F_{corr}^{1/4}$ is used in place of $F_{warp}^{1/4}$, while the next iteration still warps the right-view features using $d^{(k)}$ to obtain $\{F_{warp}^s\}$. Rectification is applied only at $1/4$ scale for efficiency; other scales keep F_{warp}^s for CVCU/UCSF. The uncertainty map U is re-estimated after each update, forming a closed-loop refinement that improves alignment reliability and mitigates error accumulation. The extra cost is marginal and mainly comes from CVCU's local correlation within $\Delta d \in [-r, r]$.

IV. Experiments

A. Datasets

We evaluate CURA-Stereo on five widely used stereo benchmarks: Scene Flow, KITTI 2012, KITTI 2015, Middlebury 2014, and ETH3D. **Scene Flow** is a large-scale synthetic dataset with dense ground-truth disparities (35,454 training and 4,370 test pairs), and we follow common practice by using the Finalpass subset. **KITTI 2012** [25] and **KITTI 2015** [26] are real-world driving benchmarks with sparse LiDAR-based annotations, containing 194/195 and 200/200 training/test pairs, respectively. **Middlebury** [27] is a high-resolution indoor dataset with 15 training and 15 test pairs, featuring challenging illumination and color variations. **ETH3D** [28] is a gray-scale benchmark with diverse indoor/outdoor scenes, comprising 27 training and 20 test pairs.

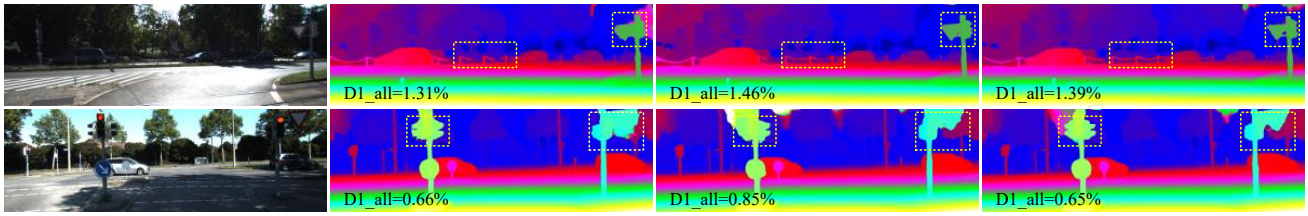
B. Implementation Details

We implement CURA-Stereo in PyTorch and conduct all experiments on NVIDIA A100-40G GPUs. For fair comparison, we follow the training protocol of IGEV-Stereo [10] unless otherwise stated. We use the AdamW [31] optimizer with gradient clipping in $[-1,1]$ and a one-cycle learning rate schedule with an initial learning rate of 2×10^{-4}. The model is pretrained on Scene Flow (Finalpass) for 200k steps with a batch size of 8 and random crops of size 320×720,

TABLE I

PERFORMANCE COMPARISON ON KITTI 2012 AND KITTI 2015 UNDER THE OFFICIAL EVALUATION METRICS. CURA-STEREO CONSISTENTLY IMPROVES ACCURACY IN CHALLENGING ILL-POSED REGIONS WHILE MAINTAINING A REASONABLE RUNTIME. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND UNDERLINED, RESPECTIVELY. ALL SUBSEQUENT TABLES ADOPT THE SAME NOTATION.

Method	KITTI 2012 [25]						KITTI 2015 [26]			Run-time (s)
	2-noc	2-all	3-noc	3-all	EPE-noc	EPE-all	D1-bg	D1-fg	D1-all	
ACVNet [21]	1.83	2.35	1.13	1.47	0.4	0.5	1.37	3.07	1.65	0.20
RAFT-Stereo [9]	1.92	2.42	1.30	1.66	0.4	0.5	1.58	3.05	1.82	0.38
PCWNet [8]	1.69	2.18	1.04	1.37	0.4	0.5	1.37	3.16	1.67	0.38
CREStereo [13]	1.72	2.18	1.14	1.46	0.4	0.5	1.45	2.86	1.69	0.41
IGEV-Stereo [10]	1.71	2.17	1.12	1.44	0.4	0.4	1.38	2.67	1.59	0.18
Selective-IGEV [3]	1.59	**2.05**	1.07	1.38	0.4	0.4	1.33	2.61	1.55	0.24
BANet-3D [29]	2.08	2.71	1.27	1.72	0.5	0.5	1.52	3.02	1.77	**0.03**
WCG-Net [15]	-	-	**1.04**	1.38	0.4	0.4	1.35	2.37	**1.52**	-
RobuSTereo [30]	-	-	-	-	-	-	1.43	2.67	1.64	0.20
CURA-Stereo(Ours)	**1.56**	2.08	1.06	**1.36**	0.4	0.4	**1.32**	**2.36**	**1.52**	0.28



(a) Left Image (b) IGEV-Stereo (c) Selective-Stereo (d) CURA-Stereo(Ours)

Fig. 5. Qualitative results on the test set of KITTI 2015. CURA-Stereo produces fewer outliers and sharper boundaries compared to IGEV-Stereo and Selective-Stereo, especially in detailed and weak-texture regions (highlighted by dashed boxes).

and the refinement module is unrolled for 22 iterations during training. We set λ_c and λ_r in Eq. (15) via grid search over $\lambda_c \in \{0.05, 0.1, 0.2, 0.3\}$ and $\lambda_r \in \{0, 0.001, 0.01, 0.1\}$, and adopt the best-performing configuration $\lambda_c = 0.2$ and $\lambda_r = 0.01$ for all experiments.

C. Comparison with State-of-the-Art

KITTI 2012 and KITTI 2015. As reported in Table I, CURA-Stereo achieves strong and highly competitive performance across most metrics on both benchmarks. On KITTI 2012, it improves IGEV-Stereo from 1.71 to 1.56 on 2-noc and from 1.44 to 1.36 on 3-all. On KITTI 2015, it achieves the best D1-fg of 2.36 (vs. 2.67 by IGEV-Stereo) and reduces D1-bg from 1.38 to 1.32, indicating more robust refinement in challenging regions. Fig. 5 presents qualitative comparisons on the KITTI 2015 test set, where CURA-Stereo produces fewer outliers and sharper boundary transitions in detailed and weak-texture areas. In addition, Fig. 6 visualizes the error-map comparisons on the KITTI 2012 test set, further showing that our method effectively suppresses large errors around object boundaries and ambiguous regions. These qualitative observations are consistent with the quantitative improvements, supporting CURA-Stereo's effectiveness in mitigating error accumulation in ill-posed areas.

Scene Flow. As shown in Table II, CURA-Stereo achieves the lowest EPE of 0.44 px on the Scene Flow test set, outperforming IGEV-Stereo (0.47 px) and IVF-AStereo (0.53 px). Table III further reports complementary metrics, where CURA-Stereo obtains the best AvgErr of 0.44 px and the lowest

Fig. 6. Error-map comparisons on the KITTI 2012 test set. Black indicates accurate regions and white denotes large disparity errors. Red boxes highlight challenging areas such as object boundaries and weak-texture regions.

>1px error rate of 4.88%, indicating more accurate and stable refinement for small-to-medium disparity errors. Although the >3px metric is not the lowest, our method achieves competitive performance, suggesting consistent convergence on large-scale synthetic data.

ETH3D. We evaluate cross-domain robustness on ETH3D under the ETH3D-all setting. As shown in Table IV, CURA-Stereo achieves the Bad 2.0 of 0.32, improving over strong baselines such as IGEV-Stereo (0.54%), Selective-IGEV (0.51%), and WCG-Net (0.36%). It also yields competitive Bad 1.0/Bad 4.0 and a low AvgErr of 0.16 px, indicating stable generalization under domain shifts. As shown in Fig. 1, CURA-Stereo further produces more coherent geometry and sharper boundaries, particularly in weak-texture regions.

TABLE II
QUANTITATIVE EVALUATION ON SCENE FLOW TEST SET.

Method	PSMNet [1]	GwcNet [2]	GANet [18]	ACVNet [21]	IGEV-Stereo [10]	IVF-AStereo [32]	BANet-3D [29]	CURA-Stereo(Ours)
EPE(px)	1.09	0.76	0.84	0.48	0.47	0.53	0.51	**0.44**

TABLE III
QUANTITATIVE ANALYSIS ON THE SCENE FLOW DATASET.

Method	>3px (%)	>1px (%)	AvgErr(px)
RAFT-Stereo [9]	3.73	6.53	0.53
PCW-Net [8]	4.23	8.06	0.78
IGEV-Stereo [10]	2.47	5.21	0.47
IVF-AStereo [32]	2.59	-	0.53
BANet-3D [29]	**2.21**	-	0.51
WCG-Net [15]	2.32	4.99	0.45
CURA-Stereo(Ours)	2.38	**4.88**	**0.44**

TABLE IV
QUANTITATIVE COMPARISON ON THE ETH3D BENCHMARK UNDER THE ETH3D-ALL SETTING.

Method	ETH3D-all			
	Bad 1.0	Bad 2.0	Bad 4.0	AvgErr(px)
GANet [18]	2.40	0.62	0.20	0.22
IGEV-Stereo [10]	1.51	0.54	0.41	0.20
RAFT-Stereo [9]	2.60	0.56	0.22	0.19
MC-Stereo [33]	1.87	0.57	0.38	0.18
Selective-IGEV [3]	1.56	0.51	0.28	**0.15**
WCG-Net [15]	**1.42**	0.36	**0.19**	**0.15**
CURA-Stereo(Ours)	1.48	**0.32**	0.20	0.16

TABLE V
OVERALL ABLATION STUDY OF KEY COMPONENTS IN CURA-STEREO ON THE SCENE FLOW TEST SET. CVCU ESTIMATES THE UNIFIED UNCERTAINTY MAP U, WHICH FURTHER GUIDES UCSF AND UDAR.

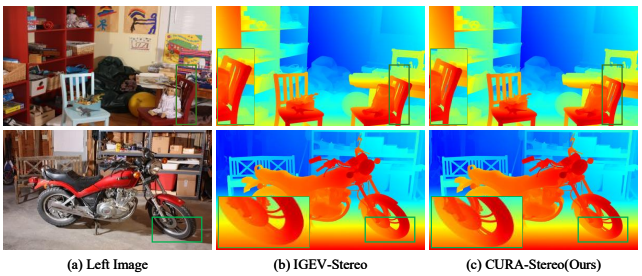
Module ↓			Metric ↓		
CVCU	UCSF	UDAR	>1px (%)	>3px (%)	Avg (px)
✗	✗	✗	5.21	2.47	0.47
✔	✔	✗	5.07	2.42	0.47
✔	✗	✔	5.09	**2.37**	0.46
✔	✔	✔	**4.88**	2.38	**0.44**

TABLE VI
SYNTHETIC-TO-REAL GENERALIZATION EXPERIMENTS WITHOUT FINE-TUNING. BAD 2.0 (%) IS REPORTED ON ETH3D AND MIDDLEBURY UNDER HALF/QUARTER RESOLUTIONS.

Method	ETH3D	Middlebury	
		half	quarter
GANet [18]	6.5	20.3	11.2
CFNet [34]	5.8	15.8	9.8
Selective-IGEV [3]	5.4	9.2	6.5
IGEV-Stereo [10]	3.6	**7.2**	6.2
RAFT-Stereo [9]	3.2	8.9	7.6
GREAT [35]	3.8	8.6	5.1
CURA-Stereo(Ours)	**3.1**	8.1	**5.0**

D. Ablation Study

Table V summarizes the ablation results of CURA-Stereo on Scene Flow. Adding CVCU+UCSF reduces the >1px error from 5.21% to 5.07% with comparable AvgErr (0.47 px), suggesting more stable refinement in ambiguous regions. In contrast, CVCU+UDAR mainly suppresses large-error tails, achieving the lowest >3px error (2.37%) and reducing AvgErr to 0.46 px, validating the benefit of uncertainty-guided rectification. With all components enabled, CURA-Stereo achieves the best overall performance (4.88% >1px and 0.44 px AvgErr) while remaining competitive on >3px. Overall, UCSF improves robustness for small-to-medium errors, whereas UDAR is more effective for severe misalignment; together they form a closed-loop uncertainty-guided refinement.



(a) Left Image (b) IGEV-Stereo (c) CURA-Stereo(Ours)

Fig. 7. Qualitative generalization results on the Middlebury training set without fine-tuning (Scene Flow → Middlebury). We compare IGEV-Stereo and CURA-Stereo.

E. Synthetic-to-Real Generalization

Table VI reports synthetic-to-real generalization results, where models are trained on Scene Flow and evaluated on real-world benchmarks without fine-tuning. We report the Bad 2.0% metric on ETH3D and Middlebury under half/quarter resolutions. CURA-Stereo achieves the best performance on ETH3D (3.1), outperforming IGEV-Stereo (3.6), RAFT-Stereo (3.2), and GREAT (3.8). On Middlebury, CURA-Stereo achieves competitive performance at half resolution (8.1) and obtains the best result at quarter resolution (5.0), improving over GREAT (5.1) and IGEV-Stereo (6.2). Overall, these results demonstrate that the proposed uncertainty-guided refinement generalizes favorably under domain shifts. Fig. 7 further provides qualitative comparisons on the Middlebury training set, where CURA-Stereo yields more coherent geometry and sharper boundaries in challenging regions.

V. CONCLUSION

We propose CURA-Stereo, an uncertainty-driven iterative stereo framework that estimates a unified uncertainty map from multi-scale cross-volume inconsistency to guide uncertainty-conditioned feature fusion and alignment rectification, thereby suppressing warping-induced error accumulation in ill-posed regions. Experiments on Scene Flow, KITTI 2012/2015, ETH3D, and Middlebury demonstrate consistent improvements over strong iterative baselines with negligible overhead.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants U25A20480, U2441241, 62272209, 62401243, and 62222206; in part by the Major Research and Development Project of Jiangxi Province under Grants 20232ACC01007, 20242BAB20038, and 20253BAC280056; in part by the Key Research and Development Program of Jiangxi Province under Grant 20232BBE50006; in part by the Early-Career Young Scientists and Technologists Project of Jiangxi Province under Grants 20244BCE52116 and 20244BCE52117; and in part by the Natural Science Foundation of Jiangxi Province under Grants 20242BAB20048 and 20252BAC240010.

REFERENCES

[1] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5410–5418.

[2] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3273–3282.

[3] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 701–19 710.

[4] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1, pp. 7–42, 2002.

[5] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *2011 IEEE intelligent vehicles symposium (IV)*. Ieee, 2011, pp. 963–968.

[6] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2722–2730.

[7] G. N. DeSouza and A. C. Kak, "Vision for mobile robot navigation: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 2, pp. 237–267, 2002.

[8] Z. Shen, Y. Dai, X. Song, Z. Rao, D. Zhou, and L. Zhang, "Pcw-net: Pyramid combination and warping cost volume for stereo matching," in *European conference on computer vision*. Springer, 2022, pp. 280–297.

[9] L. Lipson, Z. Teed, and J. Deng, "Raft-stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[10] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 919–21 928.

[11] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[12] H. Zhao, H. Zhou, Y. Zhang, J. Chen, Y. Yang, and Y. Zhao, "High-frequency stereo matching network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1327–1336.

[13] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 263–16 272.

[14] J. Jing, J. Li, P. Xiong, J. Liu, S. Liu, Y. Guo, X. Deng, M. Xu, L. Jiang, and L. Sigal, "Uncertainty guided adaptive warping for robust and efficient stereo matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3318–3327.

[15] Y. Hong, C. He, Z. Rao, Z. Chen, N. Li, and C. Zhang, "Wcg-net: Warping consistency compensation guided multi-feature fusion for stereo matching," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.

[16] J. Cheng, X. Yang, Y. Pu, and P. Guo, "Region separable stereo matching," *IEEE Transactions on Multimedia*, vol. 25, pp. 4880–4893, 2022.

[17] J. Cheng, G. Xu, P. Guo, and X. Yang, "Coatrsnet: Fully exploiting convolution and attention for stereo matching by region separation," *International Journal of Computer Vision*, vol. 132, no. 1, pp. 56–73, 2024.

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

[19] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. 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*, 2016, pp. 4040–4048.

[20] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 66–75.

[21] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 981–12 990.

[22] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1592–1599.

[23] P. Weinzaepfel, T. Lucas, V. Leroy, Y. Cabon, V. Arora, R. Brégier, G. Csurka, L. Antsfeld, B. Chidlovskii, and J. Revaud, "Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 969–17 980.

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[26] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3061–3070.

[27] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German conference on pattern recognition*. Springer, 2014, pp. 31–42.

[28] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. 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*, 2017, pp. 3260–3269.

[29] G. Xu, J. Liu, X. Wang, J. Cheng, Y. Deng, J. Zang, Y. Chen, and X. Yang, "Banet: Bilateral aggregation network for mobile stereo matching," *arXiv preprint arXiv:2503.03259*, 2025.

[30] Y. Wang, Y. Liang, Y. Hu, and Y. Fu, "Robustereo: Robust zero-shot stereo matching under adverse weather," *arXiv preprint arXiv:2507.01653*, 2025.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[32] Y. Gao and L. Shen, "Iterative volume fusion for asymmetric stereo matching," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 10 907–10 914.

[33] M. Feng, J. Cheng, H. Jia, L. Liu, G. Xu, Q. Hu, and X. Yang, "Mc-stereo: Multi-peak lookup and cascade search range for stereo matching," *arXiv preprint arXiv:2311.02340*, 2023.

[34] Z. Shen, Y. Dai, and Z. Rao, "Cfnet: Cascade and fused cost volume for robust stereo matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 906–13 915.

[35] J. Li, X. Chen, Z. Jiang, Q. Zhou, Y.-H. Li, and J. Wang, "Global regulation and excitation via attention tuning for stereo matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 25 539–25 549.