

Edge-Guided Geometry Attention for Iterative Stereo Matching

Nan Zhang

Qingdao University of Technology

Yuqing Liu

923925985@qq.com

Qingdao University of Technology

Jing Li

Qingdao University of Technology

Kunzhen Wu

Qingdao University of Technology

Research Article

Keywords: Stereo Matching, Cost-Volume Fusion, Edge-Guided, Attention Mechanism, Deep Learning

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11189696/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Edge-Guided Geometry Attention for Iterative Stereo Matching

Abstract

Iterative stereo matching methods combine compact geometric representations with recurrent disparity refinement, but ambiguous correspondences, information loss during volume compression, and boundary over-smoothing remain challenging. We present EGA-Stereo, an iterative framework that combines Channel-wise Gated Cost Volume Fusion (CGCVF), Geometry-aware Feature Squeeze (GFS), and Edge-Guided Geometry Modulation (EGGM). CGCVF integrates raw correlation responses with aggregated geometric context, GFS performs disparity-wise attention aggregation to form compact geometric features, and EGGM injects learned edge cues before recurrent updates. On Scene Flow, EGA-Stereo achieves an End-Point Error (EPE) of 0.45 px and a >3 px error rate of 2.26%. It also obtains competitive results on KITTI and consistent zero-shot improvements over the IGEV-Stereo baseline on Middlebury and ETH3D. These results suggest that jointly modeling matching reliability, compact geometry, and boundary cues is useful for iterative stereo matching.

Keywords: Stereo Matching; Cost-Volume Fusion; Edge-Guided; Attention Mechanism; Deep Learning

1. Introduction

Stereo matching estimates dense disparity from rectified image pairs and is a fundamental component of depth perception for autonomous driving, robotics, and 3D reconstruction. Learning-based stereo networks generally extract image features, construct a disparity-dependent cost volume, aggregate contextual evidence, and regress a disparity map (Chang and Chen 2018; Guo et al. 2019; Kendall et al. 2017; Zhang et al. 2019). Although this pipeline has improved prediction accuracy, dense three-dimensional aggregation remains computationally demanding. Correspondence ambiguity also persists in weakly textured, repetitive, reflective, occluded, and large-disparity regions. Local appearance may not determine a unique correspondence, whereas strong regularization may smooth the geometric variations needed to recover thin structures and depth discontinuities.

Iterative stereo matching has progressively shifted the field from one-shot disparity regression toward recurrent refinement of reusable matching representations. RAFT-Stereo (Lipson et al. 2021) introduced recurrent queries over a multilevel correlation representation, and CREStereo (Li et al. 2022) incorporated cascaded multiscale updates to improve the use of correspondence evidence. IGEV-Stereo (Xu et al. 2023) further combined local correlation lookup with an aggregated geometry encoding volume. Subsequent methods have introduced frequency-selective recurrent refinement (Wang et al. 2024), local-structure-guided disparity propagation (Li et al. 2024), and specialized representations for large disparity variations (Zhao et al. 2026). These developments indicate that recurrent accuracy depends not only on the update operator, but also on which matching and geometric cues are retained before each update.

Existing representation research can be grouped into three complementary directions. The first direction focuses on cost-volume construction and fusion, using correlation grouping, attention, adaptive weighting, or uncertainty estimation to improve the quality of matching evidence (Chen et al. 2024; Jing et al. 2023; Xu et al.

2022; Zhu et al. 2025). The second direction seeks compact geometric representations for iterative refinement through candidate-range prediction, disparity compression, adaptive downsampling, or redesigned aggregation (Guo et al. 2019; Guo et al. 2023; Wang et al. 2025; Wang et al. 2024; Wang et al. 2022). The third direction introduces structural guidance, including frequency, neighbourhood, semantic, or edge cues, to improve local propagation and disparity discontinuities (Li et al. 2024; Song et al. 2019; Wang et al. 2024). Recent zero-shot approaches further highlight the value of transferable geometric priors (Bartolomei et al. 2025), and benchmark studies show that accuracy, efficiency, and generalization should be evaluated together (Guo et al. 2023). However, these directions are often developed at separate stages, and their interaction across cost-volume fusion, geometric compression, and recurrent boundary refinement remains insufficiently characterised.

At the cost-volume fusion stage, iterative methods must balance local correspondence detail against contextual regularisation. Raw correlation volumes preserve detailed local responses, but they also contain redundant or ambiguous matches. Aggregated geometry volumes provide broader contextual structure and can reduce part of this ambiguity, yet aggregation may attenuate the local responses required to recover thin objects and depth discontinuities. Simple concatenation or direct replacement may not explicitly distinguish complementary local evidence from unreliable responses. This motivates a selective fusion mechanism that retains the contextual representation while reintroducing useful correlation cues.

A second limitation arises when a high-dimensional cost or geometry volume is converted into the compact representation used by recurrent updates. Uniform pooling or content-invariant compression applies the same disparity-reduction rule to pixels with different ambiguity patterns. Such operations may underweight or discard informative disparity responses in occluded, weakly textured, or large-disparity regions while retaining redundant responses elsewhere. The compression process therefore needs to adapt its disparity-wise weighting to local evidence while producing a compact geometric feature suitable for repeated lookup and recurrent updating.

Boundary guidance introduces a further challenge during recurrent refinement. Image gradients and appearance edges can indicate depth discontinuities, but some edges arise from texture or illumination changes and do not correspond to disparity boundaries. Treating such cues as hard constraints may create false discontinuities. Conversely, using them only during image encoding or as auxiliary supervision may limit their influence on later disparity updates. These observations motivate a soft structural signal that can modulate geometric features at the recurrent-update interface without directly imposing an edge-defined disparity structure.

To address these connected limitations, we propose EGA-Stereo, an iterative framework that coordinates three operations along the representation pathway. Channel-wise Gated Cost Volume Fusion (CGCVF) selectively reinjects projected raw correlation features into the aggregated geometry volume through a voxel-adaptive channel gate. Geometry-aware Feature Squeeze (GFS) then aggregates the resulting representation along the disparity dimension using learned disparity-wise attention to produce a compact geometric feature. Edge-Guided Geometry Modulation (EGGM) generates a soft edge gate from image gradients and appearance features. The gate is used in two complementary ways: it adjusts the temperature of the disparity-wise attention in GFS and modulates the compact geometric feature before ConvGRU updates. This staged design connects evidence fusion, geometric compression, and boundary-aware recurrent refinement at the points where matching detail, disparity information, and structural cues may otherwise be lost.

The contributions of this work are threefold. First, CGCVF introduces a selective cost-volume fusion strategy that preserves the aggregated geometry volume as the contextual representation while recovering complementary local correspondence evidence through voxel-adaptive channel gating. Second, GFS performs learned disparity-wise attention aggregation to convert the high-dimensional representation into a compact geometric feature for recurrent lookup. Third, EGGM provides a soft edge gate that is reused for edge-conditioned attention-temperature scaling and pre-update geometric modulation, thereby introducing structural guidance without treating every

image edge as a disparity discontinuity. Together, these components establish a coordinated representation pathway from local and contextual matching evidence to compact geometry and boundary-aware disparity refinement.

The quantitative comparisons in Fig. 1 summarize the resulting accuracy and iterative behavior. As shown in **Fig. 1(a)**, EGA-Stereo reports lower Scene Flow EPE than IGEV-Stereo at every listed update count: the error decreases from 0.58 versus 0.66 after one update to 0.45 versus 0.47 after 32 updates. **Fig. 1(b)** compares representative methods on KITTI 2012 using 3-all and 3-noc, for which lower values are preferable; EGA-Stereo obtains 1.34% and 1.05%, respectively, and occupies the lower-left position among the plotted methods. Together with the results on KITTI 2015, Middlebury, and ETH3D, these comparisons indicate improved performance over the selected IGEV-Stereo baseline and competitive accuracy across the evaluated benchmarks, rather than uniform superiority under every dataset and metric.

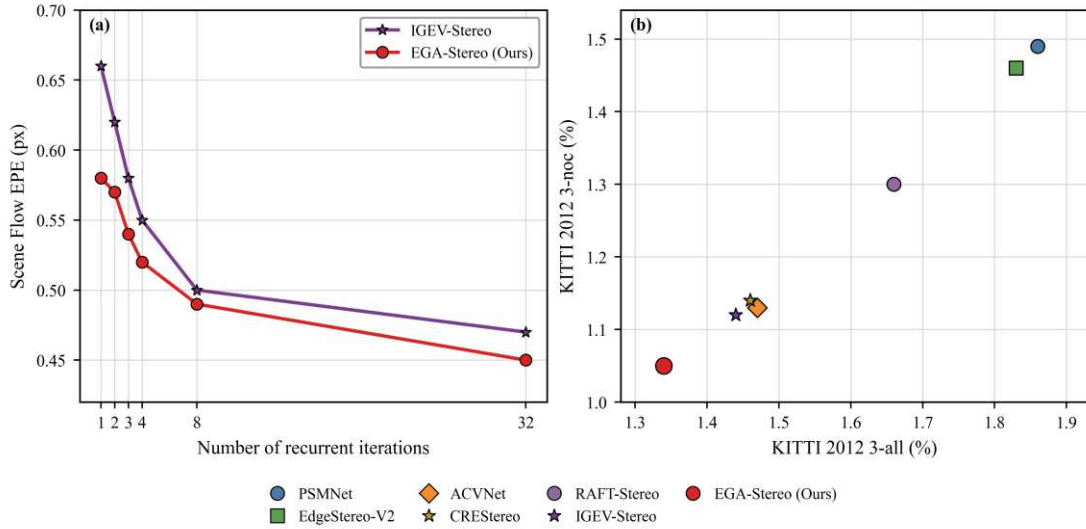


Fig. 1. (a) EPE comparison with IGEV-Stereo (Xu et al. 2023) on Scene Flow test set as the number of recurrent iterations changes. (b) Comparison with representative stereo matching methods on KITTI 2012 in terms of 3-all and 3-noc.

2. Related Work

A. Channel-wise Contextual Cost Volume Enhancement

Cost-volume construction determines the quality of correspondence evidence and the computational burden of subsequent regularization. Group-wise correlation retains multiple matching proposals with lower redundancy than full correlation (Guo et al. 2019), while ACVNet uses correlation-derived attention to suppress irrelevant content in concatenation volumes (Xu et al. 2022); both improve the informativeness of cost representations, but their channel weighting is mainly tied to the construction stage. AANet replaces expensive 3D regularization with adaptive intra-scale and cross-scale aggregation (Xu and Zhang 2020), whereas DeepPruner learns a differentiable PatchMatch module to discard unlikely disparities before dense evaluation (Duggal et al. 2019); these designs improve efficiency, but sparse sampling or early pruning may weaken hard-to-recover hypotheses around occlusions and large disparity changes. CFNet combines fused and cascaded cost volumes with uncertainty-guided search-range reduction to improve robustness across domains (Shen et al. 2021), while CVCNet predicts pixel-wise valid disparity ranges to compress the volume (Guo et al. 2023); both reduce redundant candidates, although their performance depends on reliable uncertainty or range estimation. ADStereo preserves spatial structure through adaptive downsampling and disparity-aligned upsampling (Wang et al. 2025), SPNet replaces dense local 3D aggregation with slanted-plane aggregation (Wang et al. 2022), and DCANet introduces disparity-class context for homogeneous-region aggregation (Wang et al. 2024). Overall, existing approaches improve efficiency through channel grouping, attention weighting, adaptive sampling, candidate pruning, spatial rescaling, or redesigned aggregation, yet jointly exploiting complementary raw and regularized matching evidence without suppressing weak but valid correlations remains challenging.

B. Geometry-aware Feature Compression and Extraction

Iterative optimization-based stereo methods improve efficiency by repeatedly

updating disparity from reusable geometric representations. RAFT-Stereo builds a multi-level all-pairs correlation pyramid and recurrently retrieves local matching evidence (Lipson et al. 2021), while CREStereo adopts multi-scale cascaded recurrent updates and adaptive group correlation to reduce information loss during refinement (Li et al. 2022); these designs support efficient iterative correction but can remain ambiguous in textureless, repetitive, or occluded regions because their representations contain limited non-local geometry. IGEV-Stereo complements local correlations with an aggregated geometry encoding volume and a stronger initialization (Xu et al. 2023), improving contextual reasoning and convergence, whereas HITNet represents disparity and slanted-plane hypotheses as compact tiles and hierarchically refines them without constructing a full 3D volume (Tankovich et al. 2021); the former retains richer geometry at higher cost, while the latter gains real-time efficiency but relies on a more constrained geometric representation. Selective-Stereo introduces multi-frequency branches and contextual spatial attention into the recurrent updater (Wang et al. 2024), increasing adaptivity at the expense of architectural complexity. More recently, FoundationStereo performs long-range cost-volume filtering with large-scale pretraining (Wen et al. 2025), and DEFOM-Stereo injects monocular depth-foundation features and scale updates into recurrent refinement (Jiang et al. 2025); these methods broaden contextual priors but introduce large pretrained backbones and additional adaptation or scale-consistency requirements. Across these approaches, recurrent querying progressively corrects disparity, yet compressing high-dimensional matching evidence into update-friendly geometry can discard disparity hypotheses, and balancing local detail, non-local context, memory consumption, and update efficiency remains unresolved.

C. Edge-guided Geometry Modulation and Gating

Structure-aware stereo methods use boundaries, semantics, neighborhoods, or frequency cues to address ambiguous correspondences and over-smoothed discontinuities. SegStereo embeds semantic features and semantic regularization into disparity estimation (Yang et al. 2018), while SGNet introduces semantic-guided

confidence, residual refinement, and boundary-aware smoothness constraints (Chen et al. 2021); high-level semantics improve consistency in textureless regions, but require auxiliary semantic representations and may not preserve fine geometric boundaries when category and depth edges are misaligned. StereoNet hierarchically restores high-frequency details through learned edge-aware upsampling (Khamis et al. 2018), and EdgeStereo incorporates edge features as residual-pyramid regularization (Song et al. 2019); both improve boundary localization, although image edges may include texture changes that do not correspond to disparity discontinuities. LoS models local slant, disparity offsets, and neighbor relations and propagates low-uncertainty disparities through local structures (Li et al. 2024), providing stronger geometric consistency but requiring explicit structural modeling and propagation. Selective-Stereo separates frequency information inside the recurrent updater and applies contextual spatial attention (Wang et al. 2024), whereas SMoEStereo exploits selectively adapted foundation-model experts for cross-domain robustness (Wang et al. 2025); the former increases updater complexity and the latter depends on large pretrained representations. Consequently, existing structure-aware designs improve boundaries or ambiguous regions through auxiliary semantics, image edges, local propagation, frequency decomposition, or foundation-model priors, but sensitivity to noisy guidance, semantic-depth misalignment, and the limited control of structural cues during recurrent updates remain open limitations.

3. Method

As shown in **Fig. 2**, EGA-Stereo takes a rectified stereo pair as input and extracts matching and contextual features through two dedicated networks. The matching features are used to construct a raw correlation volume, which is further aggregated to provide contextual geometric information. The raw and aggregated volumes are then fused by the proposed Channel-wise Gated Cost Volume Fusion (CGCVF) module. The fused representation is compressed by the Geometry-aware Feature Squeeze (GFS) module to generate a compact geometric feature for iterative disparity refinement. In parallel, the Edge-Guided Geometry Modulation (EGGM) module extracts structural cues from the left image and introduces edge-aware guidance into the geometric representation and recurrent update process. The resulting geometric and contextual features are processed by the ConvGRU to progressively refine the disparity estimate, which is finally restored to the original image resolution through learnable upsampling.

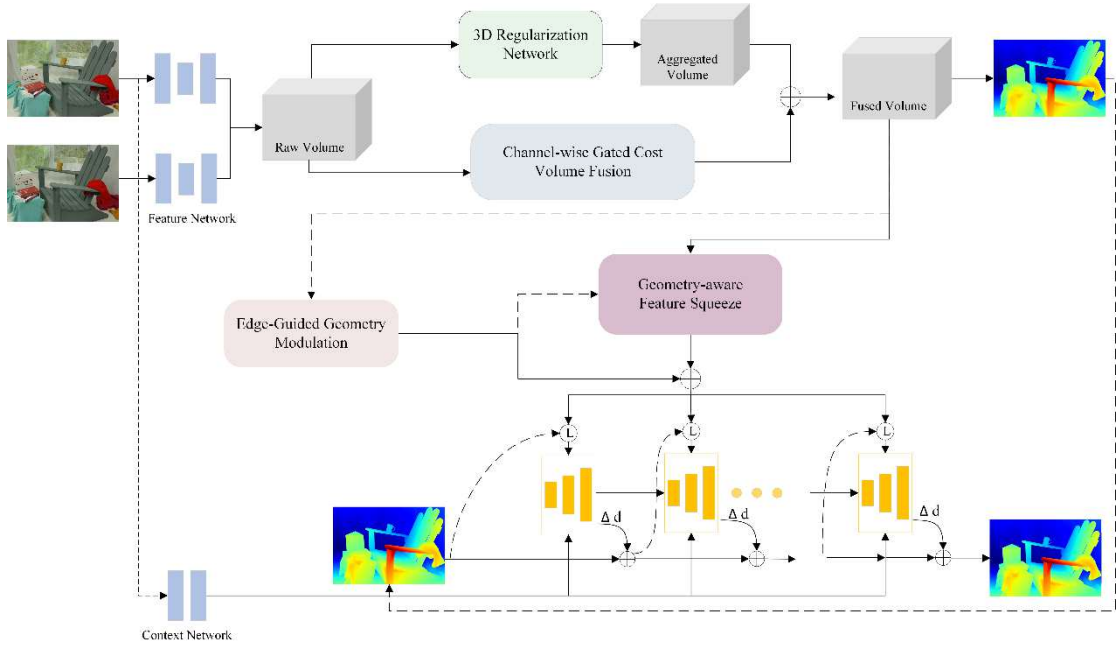


Fig. 2. Overall structure of the stereo matching network.

3.1. Multi-scale Feature Extraction

The feature extraction module consists of two distinct branches: a matching feature

network for constructing the cost volume, and a context network for driving the iterative updates of the GRU. The feature network extracts high-resolution, discriminative representations from both the left and right images to build the subsequent correlation volume. The context network, taking only the left image as input, extracts contextual features.

Feature Network: The feature network employs a weight-sharing MobileNetV2 (Sandler et al. 2018) backbone to encode both left and right images. It progressively downsamples the inputs to 1/32 resolution and utilizes deconvolutions with skip connections to upsample them back to 1/4 resolution. This outputs multi-scale features F_i , where $i \in \{4, 8, 16, 32\}$. The 1/4 resolution features are used to construct the correlation volume, while features from other scales guide the subsequent 3D cost aggregation.

Context Network: The context network processes only the left image using residual blocks to extract multi-scale contextual features at 1/4, 1/8, and 1/16 resolutions. At each scale, the features are split into two components: one initializes the ConvGRU hidden state, and the other serves as dynamic context injection during iterative updates to ensure disparity map smoothness.

3.2. Channel-wise Gated Cost Volume Fusion

Given the matching features of the left and right images at 1/4 resolution: $f_{l,4}, f_{r,4} \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$, where $C = 96$, and the search disparity range is set to $D = \frac{\max_disp}{4}$. The right-image feature is shifted along the horizontal epipolar direction, and its element-wise product with the left-image feature is computed at each disparity hypothesis. The resulting raw correlation volume is defined as

$$C_{raw}(c, d, x, y) = \begin{cases} f_{l,4}(c, y, x) f_{r,4}(c, y, x - d), & x \geq d, \\ 0, & x < d, \end{cases} \quad d \in [0, D - 1], \quad (1)$$

This yields the raw matching cost volume $C_{raw} \in \mathbb{R}^{C \times D \times \frac{H}{4} \times \frac{W}{4}}$. To establish global contextual constraints across spatial and disparity dimensions, we employ a 3D

Hourglass network (Chang and Chen 2018) to perform multi-scale cost aggregation, yielding a compact geometric context volume $C_{agg} \in \mathbb{R}^{C_a \times D \times \frac{H}{4} \times \frac{W}{4}}$, where the number of output channels is $C_a = 8$.

During 3D hourglass aggregation, intermediate cost-volume features are modulated by the corresponding left-image features through the feature-guided attention operations used in the aggregation backbone. For an intermediate cost-volume feature C_i and the corresponding left-image feature $f_{l,i}$, the feature-guided modulation is defined as

$$w_i = \sigma(g(f_{l,i})), \quad C'_i = w_i \odot C_i \quad (2)$$

Here, $g(\cdot)$ denotes the feature mapping function, $\sigma(\cdot)$ is the sigmoid normalization function, and $\odot$ denotes element-wise multiplication. The modulated cost volumes C'_i continue to participate in subsequent cross-scale feature propagation. This mechanism adaptively adjusts the response intensity of different channels based on image content, thereby enhancing discriminative features and suppressing noise interference in regions with weak texture.

Relying solely on the aggregated volume C_{agg} may weaken fine-grained correspondence responses around thin structures, occlusions, and large disparity discontinuities. To address this issue, we propose a Channel-wise Gated Cost Volume Fusion (CGCVF) module, whose architecture is illustrated in **Fig. 3**. CGCVF preserves the aggregated geometry volume as the main representation and selectively injects reliable local information from the raw correlation volume through a learned voxel-adaptive channel-wise gate.

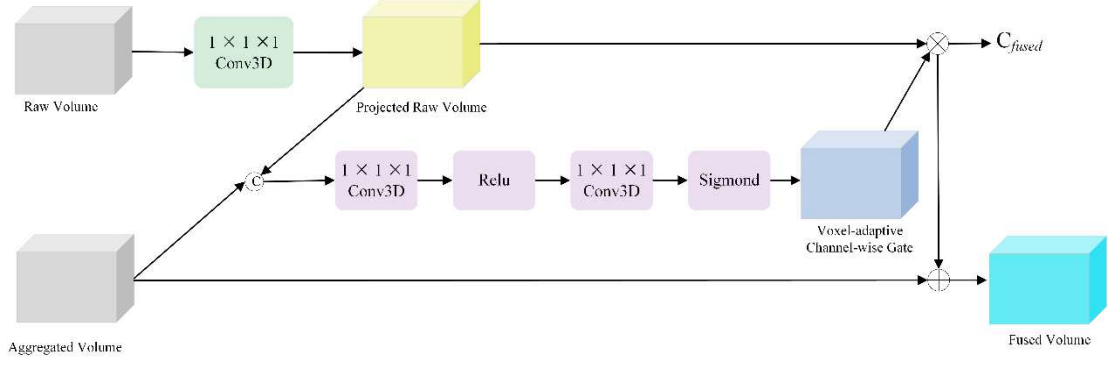


Fig. 3. Detailed architecture of the proposed Channel-wise Gated Cost Volume Fusion (CGCVF) module

First, a $1 \times 1 \times 1$ 3D convolution is applied to C_{raw} to perform channel compression and alignment, yielding the low-dimensional projected features $C_{\text{proj}} \in \mathbb{R}^{C_a \times D \times \frac{H}{4} \times \frac{W}{4}}$. Next, C_{proj} is concatenated with the aggregation volume C_{agg} along the channel dimension, and an adaptive gating weight mask G for feature selection is computed via a gating generation network composed of multiple 3D convolutional layers:

$$G = \sigma(\text{Conv}_{3D}(\delta(\text{Conv}_{3D}([C_{\text{proj}}, C_{\text{agg}}]))) \quad (3)$$

where $[\cdot, \cdot]$ denotes concatenation along the channel dimension, and $\delta(\cdot)$ represents the ReLU activation function. Guided by this gating mask, the network can adaptively evaluate the reliability of the original matching features. Finally, by injecting local details into the aggregated context through gated modulation and residual summation, the final fused feature volume is obtained:

$$C_{\text{fused}} = C_{\text{agg}} + G \odot C_{\text{proj}}, C_{\text{fused}} \in \mathbb{R}^{C_a \times D \times \frac{H}{4} \times \frac{W}{4}} \quad (4)$$

3.3. Geometry-aware Feature Squeeze

After obtaining the fused feature volume C_{fused} , we propose a geometry-aware attention squeeze mechanism whose the detailed architecture of which is illustrated in **Fig. 4.**

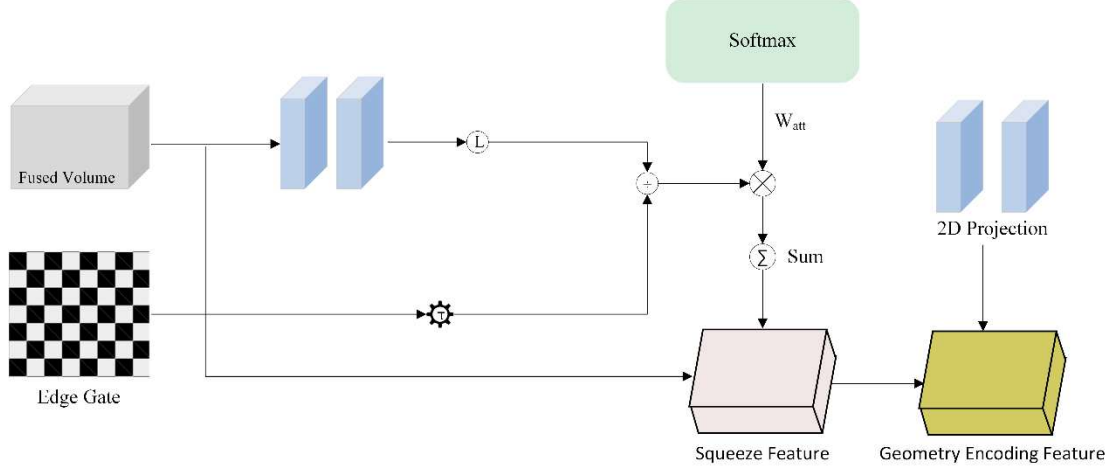


Fig. 4. Detailed architecture of the proposed Geometry-aware Feature Squeeze (GFS) module

First, a 3D classification head is introduced along the disparity dimension to perform response mapping on C_{fused} , yielding an initial disparity probability distribution. The initial disparity estimate d_0 , used to initiate the recurrent updates, is then regressed via a soft-argmin operation.

Meanwhile, to generate geometric feature maps rich in structural information, we design an edge-guided dynamic temperature attention mechanism along the disparity dimension. Specifically, we first compute the unnormalized attention logits $L_{\text{att}} \in \mathbb{R}^{1 \times D \times \frac{H}{4} \times \frac{W}{4}}$ from the cost volume using a 3D convolutional module consisting of InstanceNorm and ReLU.

We also introduce an edge-guided mask $G_{\text{edge}} \in \mathbb{R}^{1 \times \frac{H}{4} \times \frac{W}{4}}$ generated by the edge fusion module to adaptively modulate the temperature of the attention distribution spatially:

$$\begin{aligned} \tau(x, y) &= \max(1.0 - \beta \cdot G_{\text{edge}}(x, y), 0.5), \\ W_{\text{att}}(d, x, y) &= \text{Softmax}_d \left(\frac{L_{\text{att}}(d, x, y)}{\tau(x, y)} \right). \end{aligned} \quad (5)$$

where β is a learnable temperature-scaling coefficient initialized to 0.5. The edge

gate spatially modulates the Softmax temperature. When $\beta > 0$, stronger edge responses reduce τ , subject to a lower bound of 0.5, thereby sharpening the disparity-attention distribution near structural boundaries, while regions with weak edge responses retain a temperature close to 1.

After obtaining the spatially varying disparity-attention weights W_{att} , GFS compresses the fused cost volume by selectively aggregating matching evidence along the disparity dimension:

$$V_{\text{sqz}}(c, x, y) = \sum_{d=0}^{D-1} W_{\text{att}}(d, x, y) C_{\text{fused}}(c, d, x, y), \quad (6)$$

$$V_{\text{geo}} = \mathcal{P}(V_{\text{sqz}}).$$

where V_{sqz} is the disparity-compressed intermediate feature and $\mathcal{P}(\cdot)$ denotes a 2D projection block consisting of a 3×3 convolution, InstanceNorm, ReLU, and a 1×1 convolution. Rather than explicitly retaining the disparity dimension, this operation assigns greater importance to informative disparity hypotheses and integrates their evidence into the compact geometric representation V_{geo} . The resulting feature is subsequently modulated by EGGM to obtain V_{mod} before being supplied to the recurrent update module.

3.4. Edge-guided Geometry Modulation for Iterative Disparity Update

After obtaining the fused cost volume, we introduce an Edge-Guided Geometry Modulation (EGGM) module before recurrent disparity refinement. The module extracts structural cues from the left image and uses them to modulate the geometric features supplied to the ConvGRU. The overall architecture of EGGM is illustrated in **Fig. 5**.

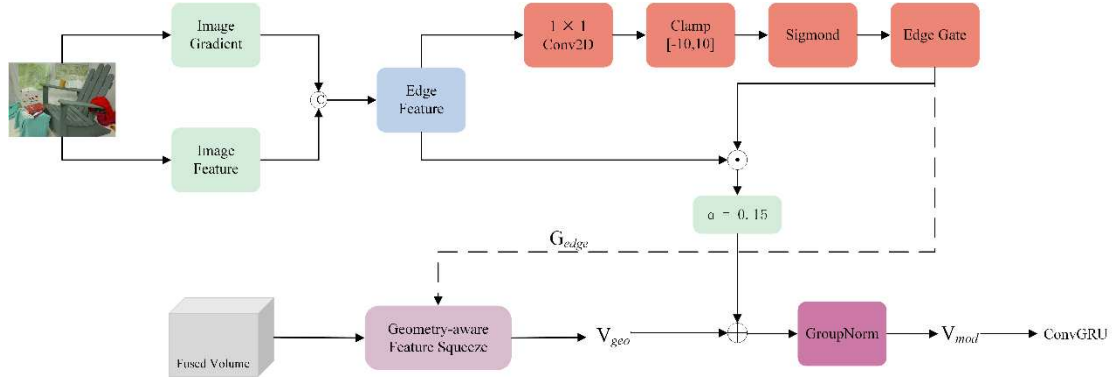


Fig. 5. Detailed architecture of the proposed Edge-Guided Geometry Modulation (EGGM) module

Given the left image I_l , we first compute absolute horizontal and vertical image differences. The differences are averaged over the RGB channels and processed by a convolutional branch to obtain gradient features. In parallel, the resized left image is processed by an image-feature branch. The two branches are concatenated and fused to produce an edge feature E_{feat} , which contains structural information associated with local image transitions.

Subsequently, a gating generator maps E_{feat} to a single-channel response using a 1×1 convolution. The response is clamped to the range $[-10, 10]$ and normalized by a sigmoid function to produce an edge gate G_{edge} . Rather than serving as a hard boundary constraint, G_{edge} provides soft spatial guidance for structurally informative regions.

The edge gate is used in two complementary ways. First, G_{edge} is provided to the Geometry-aware Feature Squeeze module as an edge mask, where it adaptively adjusts the disparity-wise attention temperature. This encourages more selective attention around structural transitions while preserving smoother responses in non-boundary regions. Second, G_{edge} spatially modulates the edge feature E_{feat} , and the resulting feature is injected into the initial geometric feature V_{geo} . Based on the obtained edge feature and edge gate, we perform spatial modulation on the initial geometric features to obtain the modulated geometric features:

$$V_{\text{mod}} = \text{Norm}(V_{\text{geo}} + \alpha \cdot (G_{\text{edge}} \odot E_{\text{feat}})) \quad (7)$$

where $\odot$ denotes element-wise multiplication, α is a learnable scaling factor initialized to 0.15, and $\text{Norm}(\cdot)$ denotes one-group GroupNorm applied to the 162-channel feature. The modulated features V_{mod} are then fed into the ConvGRU update block as geometric input, together with the multi-scale contextual features and the current disparity estimate.

At iteration k , the ConvGRU processes the previous hidden state and the current inputs through its internal update and reset gates. The recurrent update is formulated as follows:

$$\begin{aligned} z_k &= \sigma(\text{Conv}_z([h_{k-1}, x_k]) + c_z), \\ r_k &= \sigma(\text{Conv}_r([h_{k-1}, x_k]) + c_r), \\ \tilde{h}_k &= \tanh(\text{Conv}_q([r_k \odot h_{k-1}, x_k]) + c_q), \\ h_k &= (1 - z_k) \odot h_{k-1} + z_k \odot \tilde{h}_k. \end{aligned} \quad (8)$$

where z_k and r_k denote the update and reset gates, respectively, $\tilde{h}_k$ denotes the candidate hidden state, and h_k denotes the updated hidden state. The terms c_z , c_r , and c_q are context-dependent bias features generated by the context network and added to the corresponding convolutional outputs. They allow the multi-scale contextual information to modulate the gate activations and candidate-state computation during recurrent refinement.

Finally, based on the updated hidden state h_k , the current disparity residual Δd_k is predicted, and the disparity map is updated through accumulation:

$$d_{k+1} = d_k + \Delta d_k \quad (9)$$

3.5. Upsampling

After multiple rounds of ConvGRU iterations, the model obtains progressively refined disparity estimates at 1/4 resolution. To restore the original resolution and

mitigate the boundary blurring issues typically introduced by fixed interpolation methods in regions of depth discontinuity, we employ a learnable convex combination upsampling strategy. Specifically, alongside disparity prediction, a parallel convolutional branch estimates upsampling weights to describe the spatial relationship between each high-resolution pixel and its corresponding local neighborhood in the low-resolution disparity map. The high-resolution disparity is then computed as the weighted sum of the low-resolution disparities within this local neighborhood, thereby achieving content-adaptive upsampling. Finally, the upsampled disparity map from the final iteration serves as the model’s full-resolution disparity prediction.

3.6. Loss Function

During training, we employ a multi-stage supervision strategy to jointly optimize disparity estimation and edge prediction. The overall loss function consists of a disparity regression loss and an edge supervision loss, defined as:

$$L_{\text{total}} = L_{\text{disp}} + \lambda_{\text{edge}} L_{\text{edge}} \quad (10)$$

where L_{disp} represents the disparity regression loss, L_{edge} is the edge consistency constraint, and λ_{edge} is the balancing weight. To prevent the edge loss from dominating the gradients in the early stages and destabilizing the main disparity estimation task, λ_{edge} is designed as a dynamic weight that progressively increases during the training process. This progressive warm-up mechanism ensures stable convergence of the initial disparity in the early stages, while providing increasingly strong structural guidance for refining boundaries in the later stages.

Disparity Regression Loss: This term supervises the full-resolution disparity sequence $\{d_1, d_2, \dots, d_N\}$ generated during the GRU iterative process. The initial disparity is constrained by a Smooth L1 loss, while the predictions at the iterative stages are progressively supervised using a weighted L1 loss:

$$L_{\text{disp}} = L_{\text{SmoothL1}}(d_0, d_{\text{gt}}) + \sum_{i=1}^N \gamma^{\frac{15(N-i)}{N-1}} \|d_i - d_{\text{gt}}\|_1. \quad (11)$$

where $\gamma \in (0,1)$ is the decay base, N is the number of recurrent predictions, and the normalized exponent $\frac{15(N-i)}{N-1}$ assigns progressively larger weights to later predictions.

Edge Supervision Loss: To characterize disparity discontinuities, the binary ground-truth edge map E_{gt} is constructed by thresholding the sum of the absolute horizontal and vertical differences of the ground-truth disparity. The edge branch produces a clipped pre-sigmoid logit s_{edge} . Its sigmoid transformation G_{edge} is used as the soft edge gate, whereas s_{edge} is directly supervised using the weighted BCE-with-logits loss:

$$s_{\text{edge}} = \text{clip}(g_{\text{edge}}, -10, 10),$$

$$G_{\text{edge}} = \sigma(s_{\text{edge}}), \quad (12)$$

$$L_{\text{edge}} = \text{BCEWithLogits}(s_{\text{edge}}, E_{\text{gt}}; \text{pos_weight} = w_+).$$

where w_+ denotes the positive-class weight, set to 5.0, which increases the contribution of edge pixels in the BCE-with-logits loss and mitigates the class imbalance between edge and non-edge pixels.

4. Experiments

4.1 Datasets

Scene Flow (Mayer et al. 2016) is a large-scale synthetic stereo matching dataset comprising approximately 35,000 pairs of training images. The dataset provides dense and precise disparity annotations, making it suitable for the pre-training phase of models. In this experiment, we use the Finalpass subset of the Scene Flow training dataset for model pre-training.

KITTI 2012 (Geiger et al. 2012) and KITTI 2015 (Menze and Geiger 2015) are real-world stereo matching datasets designed for autonomous driving scenarios. KITTI 2012 provides 194 training pairs and 195 test pairs, while KITTI 2015 provides 200 training pairs and 200 test pairs. Middlebury 2014 (Scharstein et al. 2014) is primarily designed for high-precision stereo matching evaluation and includes a variety of indoor static scenes, comprising 15 training pairs and 15 test pairs. ETH3D (Schöps et al. 2017) is a real-world high-precision stereo dataset covering indoor and outdoor static scenes, consisting of 27 training pairs and 20 test grayscale pairs.

4.2 Implementation Details

We implement EGA-Stereo in PyTorch and conduct all experiments on an NVIDIA RTX 4090 GPU. The network is optimized using AdamW with a weight decay of 1×10^{-5} and a cosine One-Cycle learning-rate schedule with a learning rate of $2e-4$. The batch size is set to 8, and the input images are randomly cropped to 320×736 pixels using the same data augmentation strategy as IGEV-Stereo (Xu et al. 2023). We train the model on Scene Flow for 200k steps and subsequently fine-tune it on the combined KITTI 2012 and KITTI 2015 training sets for 50k steps. The maximum disparity is set to 192 pixels, and the correlation pyramid contains two levels with a lookup radius of 4. We use 22 recurrent updates during training and 32 updates during evaluation. The decay factor of the iterative disparity loss is set to 0.95. For edge supervision, the disparity-edge threshold and positive-class weight are set to 1.5 pixels and 5.0, respectively. The gradient norm is clipped to 1.0 during optimization.

4.3 Ablation Study

To quantify the contribution of each proposed component, we conduct targeted ablation experiments on the Scene Flow test set under a common training and evaluation protocol. The internal baseline is constructed by removing the Combined Geometry Encoding Volume (CGEV) from our implementation of IGEV-Stereo while retaining the same feature extractor, recurrent update operator, loss functions, training schedule, and evaluation settings. This baseline, which obtains an EPE of 0.51 px, is used only to measure the incremental effects of the proposed components and should not be interpreted as the complete IGEV-Stereo model. Because EGGM operates on the compact geometric feature produced by GFS, EGGM is evaluated only in configurations that retain GFS. We compare the individual effects of CGCVF and GFS, their combination, the addition of EGGM to GFS, and the complete three-module model. When GFS is retained but EGGM is disabled, disparity-wise attention uses a fixed temperature and no edge-guided feature is injected before the ConvGRU update. All internal variants are trained and evaluated using identical data splits, optimization settings, and recurrent-update counts.

Table 1. Ablation study of EGA-Stereo on Scene Flow.

Method	CGCVF	GFS	EGGM	EPE (px)	>3px (%)	Time (s)
Baseline				0.51	2.68	0.37
+ CGCVF	✓			0.49	2.57	0.38
+ GFS		✓		0.48	2.53	0.37
+ CGCVF + GFS	✓	✓		0.46	2.31	0.38
+ GFS + EGGM		✓	✓	0.47	2.37	0.38
EGA-Stereo (Ours)	✓	✓	✓	0.45	2.26	0.38

As shown in **Table 1**, CGCVF reduces the EPE from 0.51 to 0.49 px and the >3 px error rate from 2.68% to 2.57%, indicating that gated fusion improves the reliability of local correspondence evidence. GFS provides the largest single-module improvement, reducing the EPE to 0.48 px and the >3 px error rate to 2.53%, which supports its role in preserving informative disparity hypotheses during geometric compression. Combining CGCVF and GFS further reduces the EPE to 0.46 px and the >3 px error rate to 2.31%. When EGGM is added to GFS, the EPE decreases from 0.48 to 0.47 px and the >3 px error rate decreases from 2.53% to 2.37%, demonstrating its additional benefit during recurrent refinement. Finally, the complete EGA-Stereo model achieves the best performance, with an EPE of 0.45 px and a >3 px error rate of 2.26%. These results indicate that the three modules contribute at different stages of the pipeline and provide complementary improvements under the evaluated configurations.

The proposed EGA-Stereo maintains competitive accuracy under restricted iteration settings. **Table 2** reports the End-Point Error (EPE) comparison between EGA-Stereo and IGEV-Stereo on Scene Flow under different iteration configurations. Unlike the conventional geometry encoding volume used in IGEV-Stereo, EGA-Stereo incorporates channel-aware cost-volume fusion and edge-guided geometry modulation to provide geometric and structural cues from the early recurrent updates. As shown in **Table 2**, EGA-Stereo produces lower EPE than IGEV-Stereo at each reported update count, including 1, 2, 3, 4, 8, and 32 iterations. These results demonstrate improved early-stage disparity accuracy under a fixed number of recurrent updates.

Table 2. Scene Flow EPE comparison under different recurrent update counts.

Model	Number of Iterations					
	1	2	3	4	8	32
IGEV-Stereo (Xu et al. 2023)	0.66	0.62	0.58	0.55	0.50	0.47
EGA-Stereo (Ours)	0.58	0.57	0.54	0.52	0.49	0.45

As shown in **Table 3**, we conduct a controlled boundary-region ablation to isolate

the contributions of the two functional paths in EGGM on the Scene Flow test subset. IGEV-Stereo is included only as an external reference and is not part of the internal factorial comparison. All internal variants use the same CGCVF backbone, basic GFS squeeze operation, edge-supervision setting, training schedule, number of recurrent updates, and evaluation protocol. In the basic GFS configuration, disparity-wise attention uses a fixed temperature τ_0 . When edge-conditioned temperature scaling is enabled, the edge gate G_{edge} spatially replaces the fixed temperature with the adaptive temperature defined in Eq. (5). Geometry modulation is controlled independently and determines whether the edge-guided feature $G_{\text{edge}} \odot E_{\text{feat}}$ is injected into V_{geo} before the ConvGRU update. This design allows the individual and joint effects of the two EGGM mechanisms to be evaluated under matched conditions.

Because the purpose of this experiment is to assess boundary preservation, we report disparity-edge F1, Boundary EPE, and Boundary Bad-3. Disparity-edge F1 is computed from the predicted disparity maps and the corresponding ground-truth disparity boundaries, rather than from an auxiliary edge-prediction output. Higher disparity-edge F1 indicates better boundary localization, whereas lower Boundary EPE and Boundary Bad-3 indicate more accurate disparity estimates near depth discontinuities. The boundary subset, valid-pixel masking, boundary construction, and error thresholds are kept identical for all compared models.

Table 3. Boundary-region evaluation and ablation of EGGM components on the Scene Flow dataset. Bold: Best.

Model	Edge- conditioned temperature	Geometry modulation	Edge F1	Boundary EPE (px)	Boundary Bad-3 (%)
IGEV-Stereo (Xu et al. 2023)			0.9235	1.7515	10.4010%
CGCVF + GFS			0.9265	1.6673	9.8453%

+ Edge-conditioned temperature	✓		0.9304	1.4857	8.3379%
+ Geometry modulation		✓	0.9295	1.4917	8.4932%
EGA-Stereo (Ours)	✓	✓	0.9316	1.4752	8.2154%

The internal control, consisting of CGCVF and basic GFS, obtains an Edge F1 of 0.9265, a Boundary EPE of 1.6673 px, and a Boundary Bad-3 rate of 9.8453%. Activating edge-conditioned temperature improves the Edge F1 to 0.9304 and reduces the Boundary EPE and Boundary Bad-3 rate to 1.4857 px and 8.3379%, respectively. Geometry modulation produces a comparable improvement, yielding an Edge F1 of 0.9295, a Boundary EPE of 1.4917 px, and a Boundary Bad-3 rate of 8.4932%. The edge-conditioned temperature variant provides a slightly larger single-path improvement, whereas geometry modulation contributes an additional and complementary refinement signal. Activating both mechanisms produces the best boundary performance, with an Edge F1 of 0.9316, a Boundary EPE of 1.4752 px, and a Boundary Bad-3 rate of 8.2154%. Relative to IGEV-Stereo, the complete model reduces the Boundary EPE and Boundary Bad-3 rate by 15.8% and 21.0%, respectively. These results support the complementary roles of adaptive disparity attention and edge-guided geometric modulation in boundary-aware recurrent stereo matching.

4.4. Comparisons with State-of-the-art

We compare EGA-Stereo with state-of-the-art stereo matching methods. EGA-Stereo achieves an EPE of 0.45 px on Scene Flow, which is 0.02 px lower than that of IGEV-Stereo and competitive with those of recent methods. Detailed quantitative comparisons are provided in **Table 4**.

Table 4. Quantitative evaluation on Scene Flow test set. Bold: Best.

Method	PSMNet (Chang	AANet (Xu and	LEAStereo (Cheng et	ACVNet (Xu et al.	IGEV-Stereo (Xu et al.	EGA-Stereo (Ours)
--------	------------------	------------------	------------------------	----------------------	---------------------------	----------------------

	and Chen 2018)	Zhang 2020)	al. 2020)	2022)	2023)	
EPE (px)	1.09	0.87	0.78	0.48	0.47	0.45

Fig. 6 presents a qualitative comparison of the disparity estimates produced by EGA-Stereo and IGEV-Stereo on Middlebury. EGA-Stereo better preserves fine structures and produces more reliable estimates in complex-background regions.

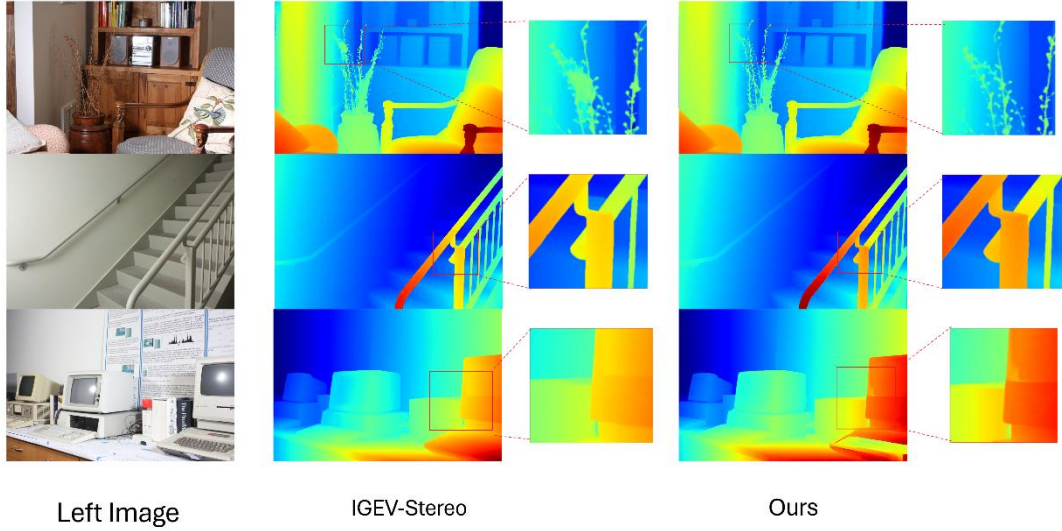


Fig. 6. Qualitative comparison on Middlebury under the zero-shot setting.

We quantitatively evaluate EGA-Stereo on the KITTI 2012 and KITTI 2015 datasets, as shown in **Table 5**. Compared to several state-of-the-art algorithms, the proposed EGA-Stereo network achieves superior performance across multiple metrics on both datasets. For the KITTI 2012 dataset, under the 3-all error metric, the proposed network achieves 1.34%, successfully outperforming IGEV-Stereo (1.44%) and RAFT-Stereo (1.66%). Similarly, on the 3-noc metric, EGA-Stereo reaches the lowest error rate of 1.05%. EGA-Stereo also improves performance in the challenging foreground regions (D1-fg), achieving an error rate of 2.53%, compared with 2.67% for IGEV-Stereo and 3.05% for RAFT-Stereo. These comprehensive results clearly demonstrate that the proposed structural optimizations effectively enhance the disparity prediction accuracy.

Table 5. Quantitative evaluation on KITTI 2012 and KITTI 2015. Bold: Best.

Method	KITTI-2012				KITTI-2015		
	3-	3-	EPE(px)	EPE(px)	D1-bg	D1-fg	D1-all
	noc(%)	all(%)	noc	all	(%)	(%)	(%)
PSMNet(Chang and Chen 2018)	1.49	1.86	0.5	0.6	1.86	4.62	2.32
EdgeStereo(Song et al. 2019)	1.46	1.83	0.4	0.5	1.84	3.30	2.08
ACVNet(Xu et al. 2022)	1.13	1.47	0.4	0.5	1.37	3.07	1.65
CREStereo(Li et al. 2022)	1.14	1.46	0.4	0.5	1.45	2.86	1.69
RAFT-Stereo (Lipson et al. 2021)	1.30	1.66	0.4	0.5	1.58	3.05	1.82
IGEV-Stereo(Xu et al. 2023)	1.12	1.44	0.4	0.4	1.38	2.67	1.59
EGA-Stereo(Ours)	1.05	1.34	0.4	0.4	1.33	2.53	1.57

To further demonstrate the superiority of the proposed EGA-Stereo in preserving depth discontinuities and recovering fine details, **Fig. 7** provides a qualitative comparison against IGEV-Stereo on the KITTI 2015 dataset. As highlighted by the red bounding boxes across various complex scenes, traditional iterative optimization in IGEV-Stereo tends to produce excessive smoothing at object boundaries and frequently fails to capture thin structures. In contrast, thanks to the explicit structural priors dynamically injected by the Edge-Guided Geometry Modulation (EGGM) module, our EGA-Stereo effectively mitigates local matching ambiguities. The visual results clearly indicate that our method is more robust in detail-rich regions and areas with complex backgrounds, successfully maintaining sharp object contours and delicate geometric structures without introducing boundary blurring.

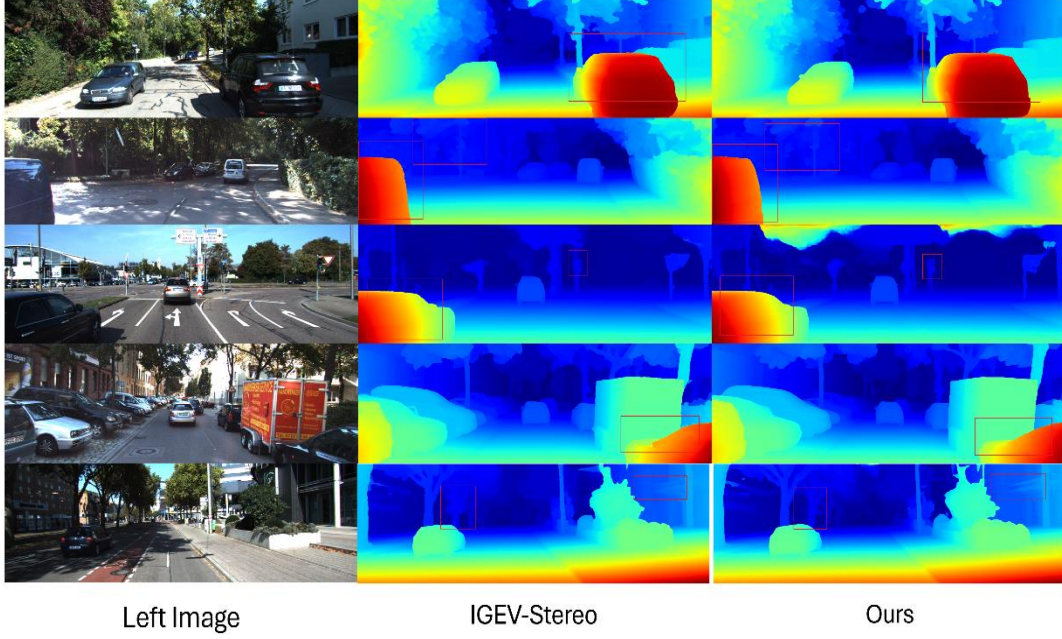


Fig. 7. Qualitative comparison results of the proposed network in the 2015 KITTI dataset.

4.5. Generalization Experiment

Real-world environments are inherently complex, featuring diverse lighting conditions, varying textures, and unpredictable object boundaries. Therefore, zero-shot generalization performance is an important indicator for evaluating the robustness of a stereo matching network. A strong generalization ability indicates that the model has learned universal geometric representations rather than dataset-specific biases.

In this section, we evaluate the cross-domain generalization performance of the proposed EGA-Stereo on unseen domains. Following standard evaluation protocols, the model is exclusively trained on the large-scale Scene Flow dataset and subsequently evaluated directly on the Middlebury 2014 and ETH3D training sets without any fine-tuning. While Scene Flow provides a massive amount of synthetic data, Middlebury and ETH3D represent complex high-resolution indoor scenes and diverse real-world scenarios, respectively, which together reflect challenging real-world conditions. As demonstrated by the quantitative results in **Table 6**, thanks to the robust geometric encoding and edge-guided structural priors, EGA-Stereo consistently improves over the

selected IGEV-Stereo baseline under the reported zero-shot protocol, while remaining competitive with the compared methods.

Table 6. Generalization experiments. All models are trained on Scene Flow. The 2-pixel error rate is used for Middlebury 2014, and the 1-pixel error rate for ETH3D.

Model	Middlebury 2014		ETH3D
	Half	Quarter	
PSMNet (Chang and Chen 2018)	25.8	14.2	23.8
GANet (Zhang et al. 2019)	13.5	8.5	6.5
STTR (Li et al. 2021)	15.5	9.7	17.2
RAFT-Stereo (Lipson et al. 2021)	8.7	7.3	3.2
IGEV-Stereo (Xu et al. 2023)	7.1	6.2	3.6
EGA-Stereo(Ours)	6.7	5.9	3.4

5. Conclusion and Future Work

This paper presented EGA-Stereo, an end-to-end edge-guided geometric attention network for iterative stereo matching. To alleviate matching ambiguities and boundary over-smoothing, EGA-Stereo incorporates three components: Channel-wise Gated Cost Volume Fusion, which combines raw correlation evidence with aggregated matching information; Geometry-aware Feature Squeeze, which constructs compact multi-scale geometric priors; and Edge-Guided Geometry Modulation, which introduces predicted structural cues into recurrent geometry updates. Experimental results on Scene Flow and KITTI demonstrate competitive stereo matching accuracy, while zero-shot evaluations on Middlebury 2014 and ETH3D show competitive cross-domain generalization.

Although EGA-Stereo achieves promising results, the multi-iteration update process still incurs computational delays that challenge real-time deployment on resource-constrained edge devices. Future work will explore lightweight recurrent structures and adaptive iteration mechanisms to further accelerate inference. Additionally, incorporating self-supervised learning strategies utilizing photometric and temporal constraints will be investigated to enhance the model's adaptability in label-scarce real-world autonomous driving scenarios.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- Bartolomei L, Tosi F, Poggi M et al (2025) Stereo Anywhere: Robust Zero-Shot Deep Stereo Matching Even Where Either Stereo or Mono Fail. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 1013-1027. <https://doi.org/10.1109/CVPR52734.2025.00103>
- Chang J-R, Chen Y-S (2018) Pyramid Stereo Matching Network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 5410-5418. <https://doi.org/10.1109/CVPR.2018.00567>
- Chen S, Xiang Z, Qiao C et al (2021) SGNet: Semantics Guided Deep Stereo Matching. In: Computer Vision – ACCV 2020. 106-122. https://doi.org/10.1007/978-3-030-69525-5_7
- Chen Z, Long W, Yao H et al (2024) MoCha-Stereo: Motif Channel Attention Network for Stereo Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 27768-27777. <https://doi.org/10.1109/CVPR52733.2024.02623>
- Cheng X, Zhong Y, Harandi M et al (2020) Hierarchical Neural Architecture Search for Deep Stereo Matching. In: Advances in Neural Information Processing Systems. 22158-22169.
- Duggal S, Wang S, Ma W-C et al (2019) DeepPruner: Learning Efficient Stereo Matching via Differentiable PatchMatch. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 4384-4393. <https://doi.org/10.1109/ICCV.2019.00448>
- Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. 3354-3361. <https://doi.org/10.1109/CVPR.2012.6248074>
- Guo X, Yang K, Yang W et al (2019) Group-wise correlation stereo network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 3273-3282. <https://doi.org/10.1109/CVPR.2019.00339>
- Guo X, Zhang C, Lu J et al (2023) OpenStereo: A Comprehensive Benchmark for Stereo Matching and Strong Baseline. arXiv preprint. 2312.00343.

<https://doi.org/10.48550/arXiv.2312.00343>

Guo Y, Wang Y, Wang L et al (2023) CVCNet: Learning Cost Volume Compression for Efficient Stereo Matching. IEEE Transactions on Multimedia. 25:7786-7799. <https://doi.org/10.1109/TMM.2022.3228169>

Jiang H, Lou Z, Ding L et al (2025) DEFOM-Stereo: Depth Foundation Model Based Stereo Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 21857-21867. <https://doi.org/10.1109/CVPR52734.2025.02036>

Jing J, Li J, Xiong P et al (2023) Uncertainty Guided Adaptive Warping for Robust and Efficient Stereo Matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 3318-3327. <https://doi.org/10.1109/ICCV51070.2023.00307>

Kendall A, Martirosyan H, Dasgupta S et al (2017) End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 66-75. <https://doi.org/10.1109/ICCV.2017.17>

Khamis S, Fanello S, Rhemann C et al (2018) StereoNet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). 596-613. https://doi.org/10.1007/978-3-030-01267-0_35

Li J, Wang P, Xiong P et al (2022) Practical stereo matching via cascaded recurrent network with adaptive correlation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 16263-16272. <https://doi.org/10.1109/CVPR52688.2022.01578>

Li K, Wang L, Zhang Y et al (2024) LoS: Local Structure-Guided Stereo Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 19746-19756. <https://doi.org/10.1109/CVPR52733.2024.01867>

Li Z, Liu X, Drenkow N et al (2021) Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 6197-6206. <https://doi.org/10.1109/ICCV48922.2021.00614>

Lipson L, Teed Z, Deng J (2021) RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In: International Conference on 3D Vision (3DV). 218-227. <https://doi.org/10.1109/3DV53792.2021.00034>

Mayer N, Ilg E, Häusser P et al (2016) 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. 4040-4048. <https://doi.org/10.1109/CVPR.2016.438>

Menze M, Geiger A (2015) Object scene flow for autonomous vehicles. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 3061-3070. <https://doi.org/10.1109/CVPR.2015.7298925>

Sandler M, Howard A, Zhu M et al (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 4510-4520. <https://doi.org/10.1109/CVPR.2018.00474>

Scharstein D, Hirschmüller H, Kitajima Y et al (2014) High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In: X. Jiang, J. Hornegger and R. Koch (ed) Pattern Recognition, Springer International Publishing, Place, pp 31-42.

Schöps T, Schönberger J L, Galliani S et al (2017) 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 (CVPR). 2538-2547. <https://doi.org/10.1109/CVPR.2017.272>

Shen Z, Dai Y, Rao Z (2021) CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 13906-13915. <https://doi.org/10.1109/CVPR46437.2021.01369>

Song X, Zhao X, Hu H et al (2019) EdgeStereo: A Context Integrated Residual Pyramid Network for Stereo Matching. In: C. V. Jawahar, H. Li, G. Mori et al. (ed) Computer Vision – ACCV 2018, Springer, Place, pp 20-35.

Tankovich V, Hane C, Zhang Y et al (2021) HITNet: Hierarchical Iterative Tile Refinement Network for Real-Time Stereo Matching. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 14362-14372. <https://doi.org/10.1109/CVPR46437.2021.01413>

Wang X, Xu G, Jia H et al (2024) Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 19701-19710. <https://doi.org/10.1109/CVPR52733.2024.01863>

Wang Y, Li K, Wang L et al (2025) ADStereo: Efficient Stereo Matching With Adaptive Downsampling and Disparity Alignment. IEEE Transactions on Image Processing. 34:1204-1218. <https://doi.org/10.1109/TIP.2025.3540282>

Wang Y, Wang L, Li K et al (2024) Cost Volume Aggregation in Stereo Matching Revisited: A Disparity Classification Perspective. IEEE Transactions on Image Processing. 33:6425-6438. <https://doi.org/10.1109/TIP.2024.3484251>

Wang Y, Wang L, Wang H et al (2022) SPNet: Learning Stereo Matching With Slanted Plane Aggregation. IEEE Robotics and Automation Letters. 7:6258-6265. <https://doi.org/10.1109/LRA.2022.3164755>

Wang Y, Wang L, Zhang C et al (2025) Learning Robust Stereo Matching in the Wild with Selective Mixture-of-Experts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 21276-21287. <https://doi.org/10.1109/ICCV51701.2025.01976>

Wen B, Trepte M, Aribido J et al (2025) FoundationStereo: Zero-Shot Stereo Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 5249-5260. <https://doi.org/10.1109/CVPR52734.2025.00495>

Xu G, Cheng J, Guo P et al (2022) Attention Concatenation Volume for Accurate and Efficient Stereo Matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 12981-12990. <https://doi.org/10.1109/CVPR52688.2022.01264>

Xu G, Wang X, Ding X et al (2023) Iterative geometry encoding volume for stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 21919-21928. <https://doi.org/10.1109/CVPR52729.2023.02099>

Xu H,Zhang J (2020) AANet: Adaptive aggregation network for efficient stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 1959-1968.
<https://doi.org/10.1109/CVPR42600.2020.00203>

Yang G, Zhao H, Shi J et al (2018) SegStereo: Exploiting Semantic Information for Disparity Estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). 660–676. https://doi.org/10.1007/978-3-030-01234-2_39

Zhang F, Prisacariu V, Yang R et al (2019) GA-Net: Guided aggregation net for end-to-end stereo matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 185-194.
<https://doi.org/10.1109/CVPR.2019.00027>

Zhao X, Wang Z, Deng Z et al (2026) Differential Feature Guidance and Compressed Cost Volume for Large Disparity Stereo Network. Neural Networks. 202:109028. <https://doi.org/10.1016/j.neunet.2026.109028>

Zhu Y, Pei S, Liu B et al (2025) AP-Net: Attention-fused volume and progressive aggregation for accurate stereo matching. Neurocomputing. 612:128685. <https://doi.org/10.1016/j.neucom.2024.128685>