Journal of Applied Computational Intelligence

eISSN: 3136-6937

Vol. 4, No. 2, April, 2026

Spherical Epipolar Cost Aggregation for ACVNet-Based Fisheye Stereo Depth Estimation

Emil Pawlak[1], Marcin Chmielewski[1], Sebastian Zalewski[2], Leon Kamiński[2,*]

[1] *Faculty of Informatics and Information Systems, Cracow University of Technology, Krakow, 31-155, Poland*

[2] *Faculty of Electrical Engineering, University of Applied Sciences in Konin, Konin, 62-500, Poland*

Received 23 December 2025; Revised 10 April 2026; Accepted 21 April 2026; Published 26 April 2026

Abstract

Fisheye stereo cameras provide a wide field of view, while their nonlinear projection complicates correspondence search and introduces spatial variation into conventional aggregation neighborhoods. This paper presents a spherical epipolar extension of an attention concatenation volume network for depth estimation directly from calibrated fisheye images. The proposed design samples candidate correspondences through calibrated rays, constructs correlation-guided attention over inverse depth, and aggregates evidence within depth-conditioned spherical epipolar neighborhoods. A validity mask excludes unsupported samples, and a residual gate controls the propagation of ambiguous costs. Evaluation uses a common candidate grid to examine overall accuracy, angular dependence, component contributions, calibration sensitivity, and computational cost. Absolute relative depth error decreases from 0.142 for the projection-adapted baseline to 0.109 for the full design, and the proportion of pixels exceeding ten percent relative error decreases from 18.6% to 13.4%. Peripheral error decreases from 0.201 to 0.139, accompanied by a latency increase from 46 to 54 ms per stereo pair. The ordered component sequence shows its largest individual step after validity gating; the contributions of spherical support remain conditional on preceding modules. Perturbation results indicate calibration sensitivity. The formulation connects attention-based matching with fisheye geometry and provides a framework for evaluating accuracy and computational tradeoffs. The analysis identifies peripheral support alignment, candidate resolution, and calibration stability as interconnected factors in wide-field stereo depth estimation. Interpretation of the reported comparisons requires documented data provenance and execution details for independent validation.

Keywords: Fisheye Stereo, Spherical Epipolar Geometry, Cost Aggregation, Attention Concatenation Volume, Depth Estimation

1. Introduction

Wide-angle camera configurations have been investigated for feature-based aerial visual odometry, including trinocular arrangements with a fisheye view [1]. Fisheye stereo-inertial systems have also been designed to process the two cameras as independent monocular views, reducing the burden of stereo feature matching [2]. Calibration research on hyper-hemispherical lenses has extended conventional camera approximations with a-central models to describe departures from single-center projection [3]. Fisheye calibration methods have incorporated constraints on the main optical axis and epipolar-line geometry to improve parameter estimation [4]. Hemisphere-based imaging representations have been used to obtain feature matches across strongly distorted fisheye images, including their peripheral regions [5]. Such representations are relevant far from the optical axis, where equal pixel displacements correspond to different angular changes. Recent omnidirectional stereo methods combine spherical sweeping with recurrent updates to estimate

* Corresponding author: kaminski.l@pk.edu.pl

http://dx.doi.org/10.68406/jaci.2026.vol4iss2no5:59-73

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

dense correspondence [6]. The resulting correspondence geometry introduces spatial variation in candidate spacing, local support and the conditioning of depth recovery. Peripheral regions can occupy a substantial proportion of the image while providing limited parallax for certain surface orientations or viewing directions. Therefore, separate assessments of image coverage and useful depth coverage are needed. These geometric properties motivate careful construction of the candidate search domain before increasing estimator capacity. Cost regularization can resolve ambiguity among available candidates, and its effectiveness depends on sampling relevant correspondence evidence. The engineering problem is to retain the contextual advantages of modern stereo networks while maintaining a consistent relationship among viewing rays, image coordinates, and recoverable radial depth.

Learning-based stereo methods have enhanced correspondence estimation by integrating image features with an organized structure of candidate matches. Pyramid cost-volume fusion methods address ambiguity across scales and improve information exchange during stereo regularization [7]. Multi-range geometry encoding volumes provide matching information at several disparity ranges for iterative stereo refinement [8]. Xu et al. [9] introduced attention concatenation volumes that derive matching-related weights from correlation evidence and apply them to concatenated features. A concentrated attention distribution can help reject geometrically unsupported hypotheses in fisheye stereo when candidate samples correspond to valid rays. Distortion-aware rectification methods learn representations of lens distortion and predict pixel-wise flow to recover corrected fisheye images [10]. Geometry-enhanced attentive stereo methods combine geometric and texture features to improve correspondence under illumination changes and weak texture [11]. Accurate camera calibration is required for both directions. Although a learned similarity function can be internally consistent, it may produce systematically misplaced candidates due to inaccuracies in the projection parameters. Furthermore, the feature volume organized by inverse depth is different from that organized by horizontal disparity, and neighboring indices have different geometric relations, particularly at the edges of the image. Therefore, the modification should indicate which components of the original stereo system are unchanged, which operations need ray-based substitution, and which comparisons isolate these modifications. These distinctions separate the effects of aggregation from changes in search resolution and training exposure.

The unresolved issue concerns the interaction between attention filtering and spatial evidence propagation under nonlinear projection. Calibrated rays determine candidate locations, while a conventional rectangular aggregation stencil combines evidence from supports whose angular separation varies across the image. Angularly regular supports can also cross a depth boundary and propagate a confident foreground match into an occluded background location. Candidate placement and support propagation are therefore treated as coupled problems. The contribution is their explicit geometric coupling within an attention-concatenation framework; correlation attention and residual fusion provide established building blocks for this integration. Calibration-based stereo transformation is employed for candidate construction; locally spherical epipolar directions and validity-aware, learned weights are used for the support of cost aggregation. Bounded-residual gates keep the original attention-weighted features when neighboring evidence is unreliable. These design choices define the hypothesis that geometric support alignment reduces peripheral matching error beyond the improvement obtained by replacing horizontal candidate sampling alone. A second hypothesis is that the cost of this alignment may be high due to increased latency for coordinate generation and irregular feature access, even with a small increase in parameters. Testing these hypotheses requires comparisons that separate candidate location from support orientation and distinguish local matching evidence from aggregated context. Angular position, radial range, and boundary visibility provide complementary ways to examine these relationships. Calibration perturbation and candidate resolution analysis also link the geometric assumptions to the computational requirements of wide-field reconstruction. These controls connect changes in aggregate error to the mechanism introduced by the architecture.

The paper develops the geometric and computational basis of fisheye depth estimation through a progression from correspondence representations and learned cost aggregation to the proposed architecture and its controlled evaluation. Section 2 reviews fisheye correspondence representations and learned cost aggregation, emphasizing the assumptions that affect a direct adaptation of the baseline. Section 3 defines the calibrated ray search, the attention weighted feature volume, and spherical epipolar aggregation, then describes the supervision and inference procedure. The formulation uses radial depth in the reference camera and explicitly handles candidates that project outside the second image. Section 4 presents an evaluation organized around comparative accuracy, component attribution, robustness, and computational behavior. It states the data partition, metric definitions, and controlled settings before examining accuracy, distributions, and computational tradeoffs. Each analysis is linked to a corresponding data figure, with consistent metric definitions and configuration labels across the text, comparison table, and plotting code. The single comparison table provides a compact view of aggregate accuracy and resource use, while the figures examine distributions that aggregate metrics can conceal. Section 5 identifies the scope of the proposed method, the limitations of the present evidence, and the research directions that follow from the analysis. The architectural illustrations use the same component names as the method description. This organization links the model specification to its geometric, statistical, and computational assessment.

2. Related Work

2.1. Fisheye Stereo Geometry and Correspondence Representations

Automatic multi-fisheye calibration has been developed using recovered human body meshes, perspective mappings, and inter-camera geometric correspondences [12]. Under the central projection approximation adopted here, a reference ray and the translation of camera centers in a calibrated stereo pair define an epipolar plane. The intersection of this plane with a unit sphere is a great circle, whose projection into a fisheye image generally forms a curve. Geometry restricts the location of correspondence; binocular visibility is determined by occlusion and camera overlap. Occlusion, field boundaries and the restricted overlap area are still separate validity constraints. Ray-based depth learning also requires an explicit choice between radial distance and optical-axis depth. This choice determines the prediction target and the interpretation of error across the wide field. The great-circle construction assumes a central camera model and synchronized views of a static scene; axial viewpoint shifts require a different projection treatment.

Resolution adaptation presents a related challenge: implicit disparity representations enable arbitrary-scale stereo prediction while recovering details lost during downsampling [13]. For fisheye matching, resolution handling also depends on angular coverage, interpolation, and sample allocation near projection boundaries. Image reprojection into a single target chart provides convenient memory access with spatially varying effective resolution. Multiple charts preserve local detail and require consistency handling across overlapping regions. Direct fisheye matching avoids an additional image resampling stage, while feature sampling occurs when candidate rays are projected into the partner view. The comparison of these methods focuses on sampling strategy and computational organization. Fair comparisons require consistent overlap masks, depth support, and feature resolution to isolate the effects of the sampling strategy.

Multi-view stereo methods have replaced static cost volumes with iterative dynamic volumes and cross-scale propagation to reduce computational demands [14]. Spacing of candidates affects both memory consumption and the resolution in various ranges. Inverse depth sampling increases the density of radial depth resolution around the camera for high-precision boundary requirements in navigation tasks. Uncertainty in calibration leads to trajectory bias and needs to be handled separately from candidate resolution. Principal-point self-calibration for vehicle-mounted fisheye cameras has been formulated through a point-oriented representation that encodes projection distortion [15]. Adapting a pretrained stereo backbone requires alignment between the search parameterization and training target, together with an explicit definition of the pixels included in evaluation. These requirements motivate evaluation by viewing angle and radial depth interval.

2.2. Learned Cost Volumes and Geometry Aware Aggregation

Disparity-classification approaches use structural priors to simplify cost-volume aggregation while preserving information for stereo estimation [16]. The effectiveness of these priors depends on how the volume encodes spatial positions and candidate indices. Horizontal disparity volumes have a regular interpretation after pinhole rectification, whereas an inverse depth volume built from fisheye projection inherits spatially varying angular sampling. Reusing the same convolutional stencil provides regular memory access while producing spatially varying geometric support. Large receptive fields can suppress isolated mismatches, yet they may also spread a dominant surface hypothesis across a narrow object or an occlusion boundary. The support definition consequently determines how the model relates spatially neighboring evidence.

Lightweight volumetric architectures such as DMCNet investigate the resource demands of contextual stereo matching [17]. Multi-branch attention fusion combines shallow and deep features during cost aggregation, while similarity-based attention improves cost-volume construction [18]. These strategies distinguish the confidence assigned to a candidate from the richer feature information needed to refine its interpretation. In fisheye matching, attention must also exclude samples outside the overlap region; validity masking prevents padded image content from entering the normalized distribution. Repeated textures and reflective surfaces can produce high confidence matches, making local agreement relevant to the weighting of neighboring costs. A spatial gate that observes local agreement can complement candidate attention, with confidence calibration evaluated separately.

Li et al. [19] proposed content-aware cost aggregation that dynamically upsamples cost volumes using inter-scale similarity guidance derived from image features. Content-adaptive upsampling provides background for support design. Constraining support orientation through calibration is a separate geometric modeling choice considered in the present method. Explicit direction constraints reduce the set of possible sampling patterns, but they can amplify systematic calibration errors. Unconstrained offsets may compensate for small errors while weakening interpretability and requiring more varied training data. The present proposal occupies a restricted middle ground, fixing support orientation through camera geometry and learning its contribution through attention and residual gating. The evaluation must isolate this choice from changes in the number of candidates, feature channels, and aggregation depth to determine whether geometric alignment explains any observed improvement.

3. Spherical Epipolar Aggregation Network

3.1. Calibrated Rays and Inverse Depth Candidates

The proposed model, denoted SECA ACVNet for spherical epipolar cost aggregation within ACVNet, accepts a synchronized pair of calibrated central fisheye images. Both cameras share a feature extractor with quarter resolution output and a feature dimension of 32. Calibration supplies a differentiable projection operator and its inverse direction map for each camera. The rigid transformation maps reference camera coordinates into partner camera coordinates, with translation expressed in meters. Radial distance along a unit reference ray is the prediction target. A fixed set of 96 inverse depth hypotheses covers radial distances from 0.5 to 30 m. Image resizing requires corresponding adjustment of the projection parameters, and projection operations use feature coordinates consistently. The following projection equation maps each depth hypothesis to its corresponding location in the partner image:

$$\mathbf{x}'_j = \pi_2\left(\frac{\mathbf{R}\mathbf{r}(\mathbf{x})}{\rho_j} + \mathbf{t}\right), \mathbf{r}(\mathbf{x}) = \pi_1^{-1}(\mathbf{x}), \|\mathbf{r}(\mathbf{x})\|_2 = 1. \quad (1)$$

Here, $\pi_1{}^{-1}$ returns the normalized reference viewing direction, π_2 projects a spatial point into the partner image, and $\mathbf{R}$ and $\mathbf{t}$ describe the calibrated transformation. The positive inverse depth hypothesis is denoted by ρ_j. Bilinear interpolation samples partner features at the projected coordinates. A binary validity tensor identifies samples within the calibrated image domain with usable viewing directions. Attention normalization and supervision operate exclusively on valid hypotheses. Reference pixels with an empty candidate set receive an unsupported output flag. Hypotheses outside the common viewing region remain invalid even when the rectangular feature grid contains padding. The validity mask is retained throughout aggregation and output generation. This mask encodes projection-domain support. Physical visibility also depends on scene geometry, so geometrically valid candidates can still correspond to occluded surfaces.

Figure 1 presents the inference workflow, connecting shared feature extraction, calibrated candidate projection, groupwise correlation, attention filtering, spherical aggregation, and radial depth regression. The architecture retains the baseline principle of using correlation evidence to modulate concatenated features. Ray based projection replaces horizontal candidate shifts, while a geometry aligned aggregation block precedes the final regularizer. The projection and tangent operations are fixed by calibration, whereas attention, support weights, and residual gates are learned. Coordinate maps can be cached when calibration, image dimensions, and the hypothesis grid remain fixed; candidate-dependent maps must retain their depth index. Feature interpolation remains part of inference, and cached coordinate storage contributes to deployment memory. Ground truth enters the network exclusively through the training supervision path.

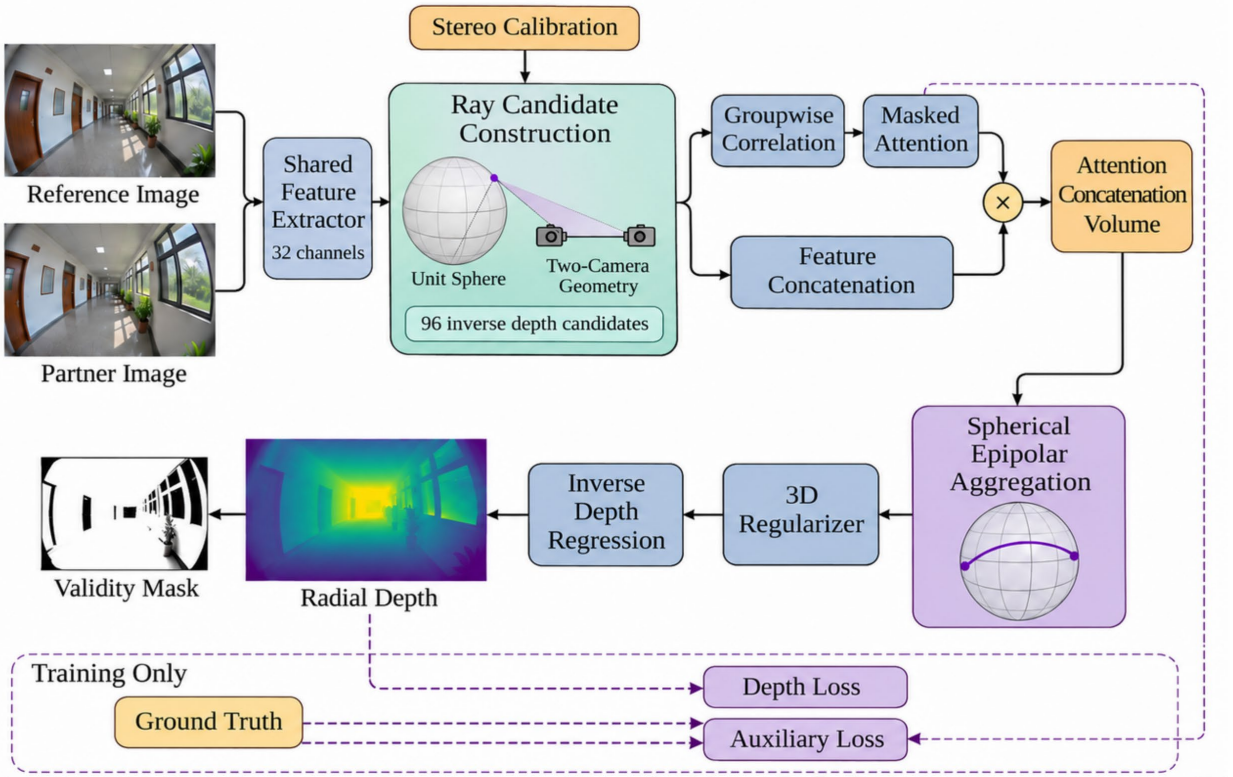


Figure 1. Calibrated fisheye stereo inference workflow

3.2. Attention Filtering and Spherical Support Aggregation

Reference and sampled partner features are divided into eight channel groups. Their normalized groupwise similarities feed a shallow attention predictor that produces one logit per candidate. A masked softmax normalizes these logits over valid inverse depth hypotheses. The resulting attention weights multiply every channel of the concatenated feature tensor. Temperature is fixed to one so that candidate resolution comparisons share the same attention normalization. In the attention weighting equation below, the binary indicator v_j determines candidate validity:

$$A_j = \frac{v_j \exp(s_j)}{\sum_\ell v_\ell \exp(s_\ell)}, \mathbf{V}_j = A_j\left[\mathbf{F}_1(\mathbf{x}), \mathbf{F}_2(\mathbf{x}'_j)\right]. \tag{2}$$

The score s_j is the predicted attention logit, and the brackets denote feature concatenation. The denominator is evaluated for reference pixels possessing at least one valid candidate. Numerical implementation subtracts the maximum valid logit before exponentiation. Empty valid sets bypass normalization and retain the unsupported-output flag, with candidate probabilities set to zero. Attention controls candidate selection, while subsequent aggregation examines agreement among neighboring representations. Repeated patterns can produce several strong modes, whereas textureless regions can yield a diffuse attention distribution. The concatenated tensor consequently preserves appearance information alongside the attention weights for later regularization. Geometry determines both candidate locations and the support used to combine neighboring costs.

At each reference direction, the epipolar plane is determined by that direction and the second camera center expressed in reference coordinates. The partner-camera center is obtained by applying the inverse reference-to-partner rigid transformation to the partner origin. The cross product of this center vector and the reference ray defines the plane normal, denoted by n after normalization. Its cross product with the reference ray defines the great-circle tangent. The spherical sampling equation expresses each offset direction as a rotation within this plane:

$$\mathbf{r}_k = \cos(\delta_k)\mathbf{r} + \sin(\delta_k)\mathbf{e}, \mathbf{e} = \frac{\mathbf{n} \times \mathbf{r}}{\|\mathbf{n} \times \mathbf{r}\|_2} \tag{3}$$

The five angular offsets are minus 1, minus 0.5, zero, 0.5, and 1 degree. Each offset ray is projected into the reference feature map, and its partner location is recomputed at the same radial depth hypothesis using the calibrated transformation. Each support point is independently projected at the central candidate radial distance. This sampling describes a local spherical shell; a shared physical surface can intersect neighboring rays at different radial distances. Learned weighting therefore remains important near slanted surfaces and depth discontinuities. Near an epipole, the camera-center/reference-ray cross-product norm is checked before either plane-normal or tangent normalization. Degenerate directions retain the center sample and record the reduced support; a reproducible implementation must specify the numerical tolerance used for this branch.

The sign of the plane normal determines the ordering of symmetric offsets while preserving the available support. Local visibility and surface continuity are handled through learned weighting. Depth discontinuities require particular care because a geometrically aligned neighborhood can still cross a surface boundary. A compact predictor therefore observes the central tensor, sampled tensors, angular offsets, and attention concentration. It produces normalized support weights and a residual gate between zero and one. Invalid support locations are excluded before normalization. Fusion follows the residual equation:

$$\tilde{\mathbf{V}}_j(\mathbf{x}) = \mathbf{V}_j(\mathbf{x}) + g_j(\mathbf{x})\left[\sum_k w_{jk}(\mathbf{x})\mathbf{V}_j(\mathbf{x}_k) - \mathbf{V}_j(\mathbf{x})\right]. \tag{4}$$

The nonnegative weights w_{jk} sum to one across valid offsets, and the center remains available whenever the original candidate is valid. A sigmoid produces the gate g_j. The update forms a convex interpolation between central and aggregated representations, bounding their relative contributions. This property constrains feature mixing; depth accuracy also depends on the correspondence evidence and the subsequent regularizer. A small gate preserves the central evidence, while a larger gate increases the influence of neighboring support. Two compact three-dimensional regularization blocks subsequently transform the refined volume into candidate costs. Figure 2 details tangent construction, great circle sampling, candidate reprojection, validity aware weighting, and residual fusion, with fixed geometric operations distinguished from trainable predictors.

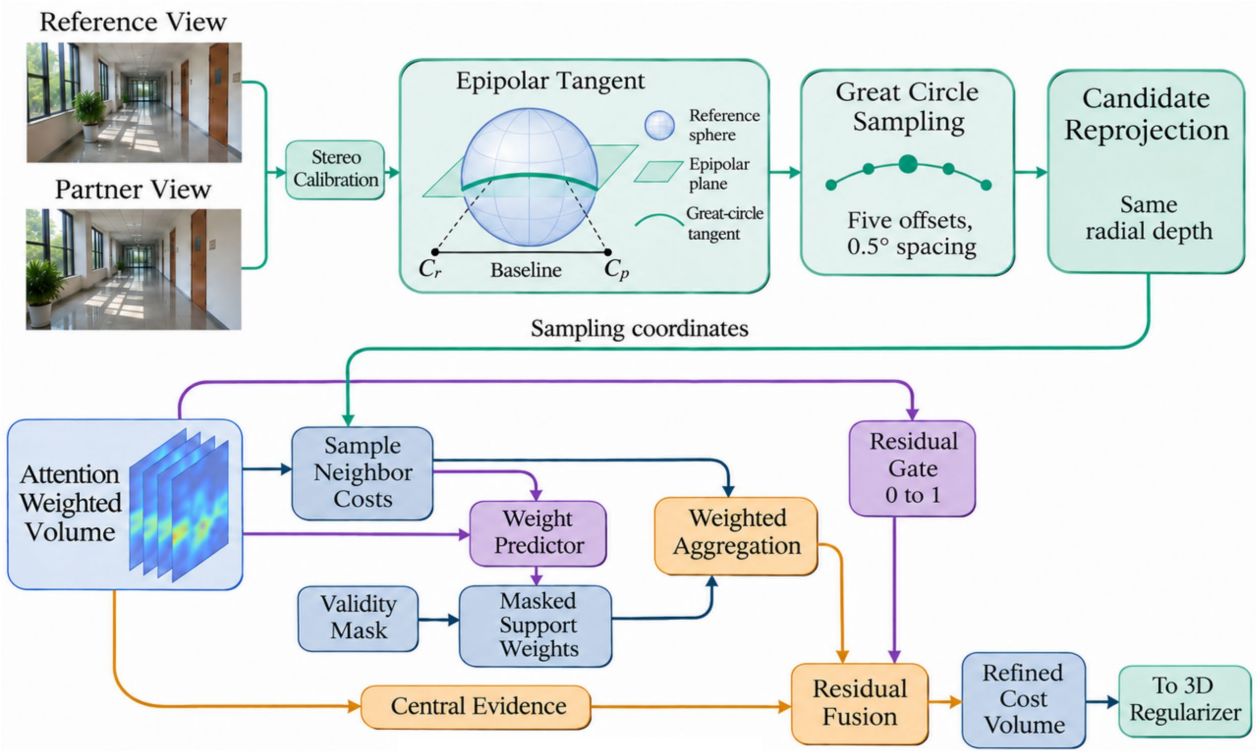


Figure 2. Spherical epipolar cost aggregation module with calibrated great circle support, validity masking, learned support weights, and bounded residual fusion

3.3. Depth Regression and Supervision

The regularizer produces a cost distribution over inverse depth hypothesis. Masked normalization restricts probability mass to valid candidates. The model regresses expected inverse depth and takes its reciprocal to obtain radial distance. This convention maintains consistency between candidate construction and prediction. Taking the reciprocal after averaging inverse depth produces a probability-weighted harmonic mean of candidate radial distances, which differs from averaging those distances directly. The depth regression equation combines candidate probabilities with the inverse depth grid:

$$P_j = \mathrm{softmax}_j(-C_j), \widehat{D} = \left(\sum_j P_j \rho_j\right)^{-1}. \quad (5)$$

Here, C_j denotes the regularized cost, and the softmax operates over valid hypotheses. The output mask identifies pixels whose candidate set is empty. The spread of the inverse-depth distribution provides a matching-dispersion indicator. Its scale depends on candidate parameterization and requires separate calibration before interpretation as uncertainty in meters. Near a multimodal boundary, expectation can produce an intermediate depth between competing surfaces, making the underlying distribution relevant to prediction assessment.

Training uses valid ground truth radial depth and a robust loss applied to the final output and an auxiliary attention prediction. The auxiliary branch follows the same inverse depth regression convention, with a loss weight of 0.3 relative to the final branch. Optimization uses AdamW with an initial learning rate of 0.0002, weight decay of 0.0001, batch size four, and a cosine schedule over 40 epochs. These settings remain consistent across architectural comparisons to control optimization and supervision. Photometric augmentation is applied independently to the image pair, while geometric augmentation updates the calibration accordingly. Cropping requires simultaneous adjustment of the projection domain and overlap masks.

The loss operates on supported rays with valid ground truth within the specified radial range. A Huber penalty with a transition at 0.2 m limits the influence of large residuals while retaining a quadratic region near correct predictions. The same supervision mask applies to all controlled variants. Angular spacing, gate width, and hypothesis count are selected using the validation partition. Inference retains the calibration and validity conventions used during training. Candidate masks and gate maps remain accessible for examining the relationship between depth predictions and geometric support. The geometric validity mask stays fixed throughout attention filtering and aggregation.

4. Experimental Evaluation and Analysis

4.1. Evaluation Protocol and Comparative Accuracy

The evaluation examines the proposed architecture under controlled geometric, photometric, and computational conditions. A common candidate grid and visibility convention allow changes in cost aggregation to be compared across the configurations. The reported partition contains 12,000 training pairs, 2,000 validation pairs, and 3,000 test pairs, with complete sequences assigned to separate partitions. The dataset release, acquisition procedure, ground-truth sensor, and sequence identifiers are unspecified in the available evaluation record. These missing records limit independent reproduction and validation of the reported values. Research on all-round stereo reconstruction shows that nominal angular coverage can include directions with very weak disparity and high reconstruction uncertainty [20]. The proposed configuration uses a 0.20 m baseline, nominal 190-degree lenses, 640 by 512 images, and a shared validity mask within a radial range of 0.5–30 m. Table 1 specifies five variants with identical candidate grids. The complete configuration changes absolute relative error from 0.142 to 0.109 and root mean square error from 1.84 to 1.43 m, while parameters increase from 6.2 to 7.1 million.

Table 1. Accuracy and computational cost under a common calibrated candidate grid

Variant	Abs Rel	RMSE (m)	Bad ten (%)	Latency (ms)	Params (M)
Ray baseline	0.142	1.84	18.6	46	6.2
Validity gate	0.132	1.72	17.0	48	6.3
Fixed sphere	0.124	1.61	15.8	50	6.5
Learned support	0.117	1.51	14.5	52	6.8
Complete	0.109	1.43	13.4	54	7.1

Absolute relative error is computed from radial depth residuals normalized by reference depth, and bad ten denotes the percentage exceeding ten percent relative error. Relative-error and bad-ten scores are first computed over the common valid pixels in each frame; frame scores then receive equal weight in the aggregate. The RMSE pooling convention requires an explicit implementation record to establish comparability with pixel-pooled benchmarks. Threshold based stereo metrics require an explicit error definition and evaluation domain [21]. Figure 3(a) presents errors of 0.142, 0.132, 0.124, 0.117, and 0.109 across the five configurations. Their progressive decline provides an ordered comparison of the added modules under constant search resolution. Figure 3(b) gives corresponding bad ten rates of 18.6%, 17.0%, 15.8%, 14.5%, and 13.4%. The complete configuration therefore changes both average deviation and threshold exceedance, with a 23.2% relative error reduction and a 5.2 percentage point decrease in bad ten.

Figure 3(c) places these configurations at latencies of 46, 48, 50, 52, and 54 ms, with marker areas encoding 6.2, 6.3, 6.5, 6.8, and 7.1 million parameters. The endpoint change corresponds to 17.4% additional latency and 14.5% additional parameters. The smaller spacing between successive error values exposes the tradeoff more clearly than an accuracy ranking alone: the final two additions consume the same four milliseconds as the preceding two additions while producing a smaller absolute error change. This pattern makes the admissible latency budget relevant to configuration selection. The stated timing protocol uses synchronized execution, batch size one, and fixed precision, with the same processing stages and input resolution. The hardware model, numerical precision value, software versions, and repetition count are unspecified, so the latency values describe the supplied configuration rather than portable device benchmarks.

Figure 3(d) gives cumulative relative error distributions constrained to the reported bad ten rates. At the ten percent threshold, cumulative coverage increases from 81.4% to 86.6%, whereas the curves converge above a relative error of 0.15. The displayed separation is concentrated near moderate errors, while the upper tails appear similar. Threshold coverage constrains one cumulative probability; residual-level observations and additional quantiles are required to establish the tail shape. Stereo-confidence methods use the response of predicted disparity to plane sweeps to assess correspondence ambiguity at individual pixels [22]. The cumulative distribution characterizes the observed error population, while pixel-level confidence would require a separate reliability assessment. Applications sensitive to rare large depth failures would require additional tail statistics. Separating threshold coverage from the mean also distinguishes changes near the decision threshold from changes concentrated in the severe error tail.

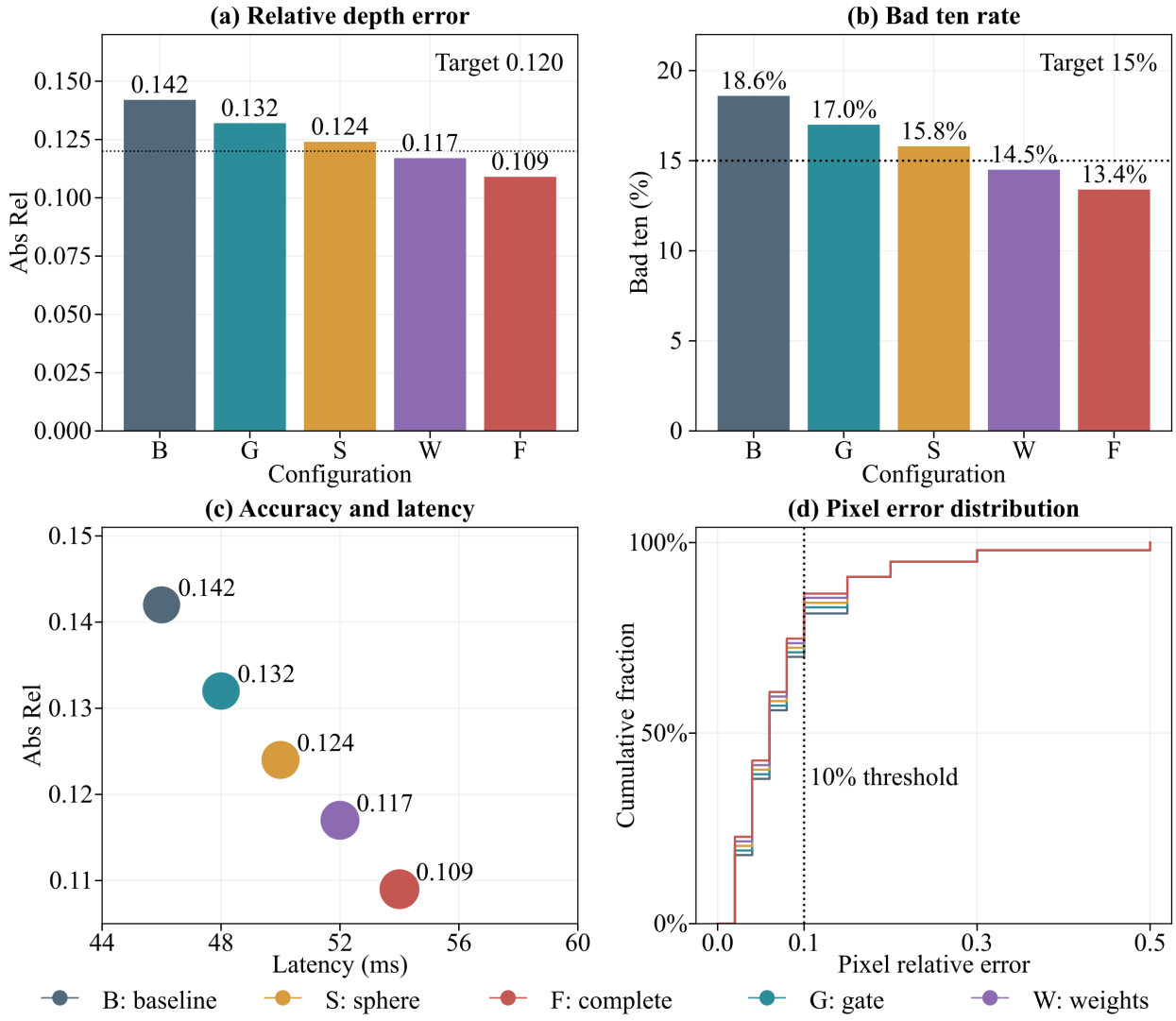


Figure 3. Comparative accuracy and resource tradeoffs. (a) Absolute relative error. (b) Bad ten rate. (c) Latency versus error with parameter scaled markers. (d) Relative error cumulative distributions

Figure 4(a) stratifies absolute relative error across six incidence bands centered at 7.5, 22.5, 37.5, 52.5, 67.5, and 82.5 degrees. Distortion-aware fisheye depth estimation explicitly incorporates projection-related information into depth prediction and refinement [23]. Baseline errors increase from 0.109 to 0.201, while complete model errors increase from 0.091 to 0.139. Incidence angle is measured from the reference optical axis. The endpoint reductions are 16.5% near the optical axis and 30.8% in the outermost band. The widening separation corresponds to the intended operating region of spherical support, where a rectangular pixel neighborhood exhibits stronger variation in angular extent. The contribution of each angular band to the image average also depends on its valid pixel count, making overlap coverage relevant to spatial comparisons.

Figure 4(b) presents frame error distributions for six radial intervals, with complete model medians of 0.075, 0.084, 0.095, 0.112, 0.138, and 0.176. Each violin contains 120 frame scores, while the central mark and thick segment identify the median and interquartile range. These displays have smaller support than the reported test partition; their selection rule, frame identities, and valid-pixel counts are needed to assess representativeness. Median error more than doubles between the nearest and farthest intervals, indicating persistent range sensitivity alongside the improvement in aggregate accuracy. The widening absolute spread also means that a single representative frame would poorly describe distant reconstruction. These interval medians describe a different aggregation level from the image averaged values in the comparison table.

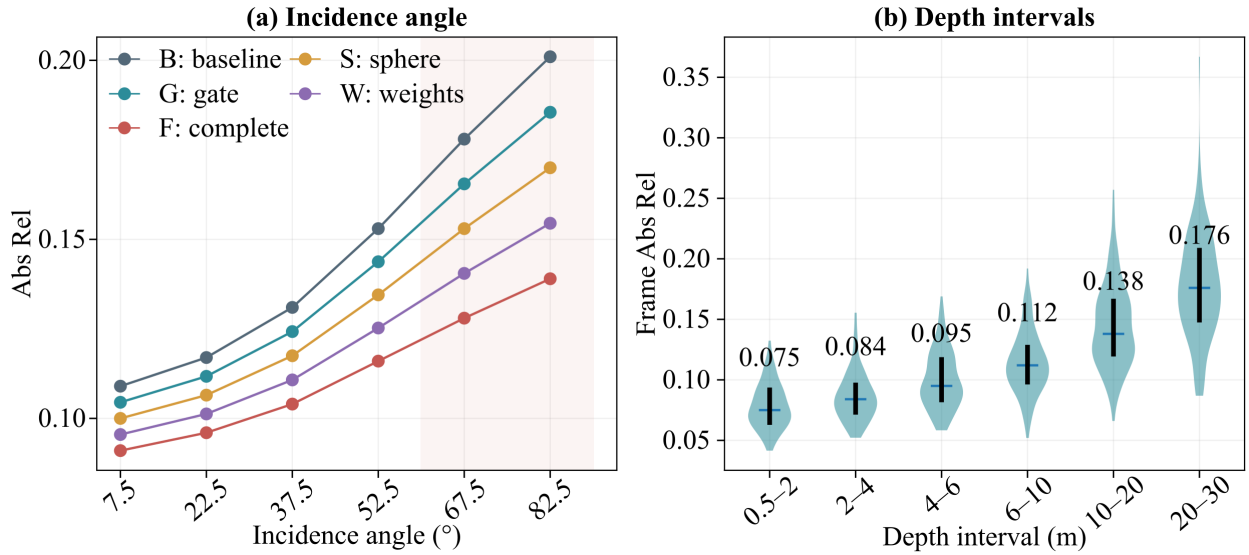


Figure 4. Spatial and depth dependent behavior. (a) Error across six incidence angle bands for five variants. (b) Frame error violin distributions across six radial depth intervals with medians and interquartile ranges

Cross stratification of incidence angle and radial range would expose whether difficult viewing directions coincide with distant surfaces. Peripheral pixels may observe nearby obstacles with strong parallax, while central pixels may correspond to distant surfaces with weak depth sensitivity. Angle and range should therefore be cross stratified while retaining the same visibility mask across configurations. Scene composition can make a peripheral band dominated by nearby surfaces appear easier to reconstruct. Boundary distance, local texture, and valid support count should accompany that stratification. Their joint examination would test whether the gate responds to ambiguous evidence at the locations were geometric support changes most strongly with viewing direction.

4.2. Component Attribution and Perturbation Behavior

Figure 5(a) follows an ordered path from 0.142 to 0.109, with successive changes of 0.010, 0.008, 0.007, and 0.008. The first step contributes 30.3% of the total reduction, while the remaining steps have similar magnitudes. Research on active disparity sampling shows that the allocation of depth guidance can affect stereo accuracy under a fixed sampling budget [24]. The sparse depth acquisition considered in that literature differs from the inverse-depth candidate grid controlled here. The candidate grid remains fixed to control hypothesis resolution. The allocation of improvement along the waterfall depends on the order in which components are introduced. The first validity-gating step is the largest individual decrease, whereas the combined later decreases include several interacting additions. Reordered or factorial comparisons would be needed to estimate independent component effects. Fixed spherical support is evaluated after the validity gate, and learned support weights are evaluated after that geometric change. The waterfall therefore specifies a reproducible sequence of comparisons through the chosen configurations, with each transition linked to an explicit architectural addition.

Figure 5(b) presents six scene conditions across all five configurations. Complete model errors are 0.091 for textured surfaces, 0.128 for weak texture, 0.146 for repeated patterns, 0.159 for reflections, 0.137 for occlusion boundaries, and 0.103 for low illumination. The corresponding baseline values are 0.118, 0.170, 0.185, 0.194, 0.183, and 0.139. Boundary error decreases by 25.1%, while reflection error decreases by 18.0%, leaving reflections as the most difficult condition. Region-aware distribution optimization has been investigated as a mechanism for improving stereo matching across structurally different image regions [25]. Edge-preserving multi-view stereo uses contextual cues and residual refinement to improve depth near boundaries and thin structures [26]. The boundary scores reported here depend on visibility labels and the chosen boundary band width, which determine how foreground transitions contribute to the aggregate metric.

Component interactions are relevant here because a validity gate changes which samples can contribute, whereas spherical support changes where those samples are obtained. Their interaction can alter the benefit of learned weighting. The scene matrix also indicates where targeted inspection would be most informative: weak-texture scores improve in the ordered comparison, while repeated patterns and reflections remain difficult. These trends motivate hypotheses about ambiguous feature evidence and support contamination. Matched attention distributions, gate activations, and support masks on the same frames would connect each condition score to the operation responsible for that behavior.

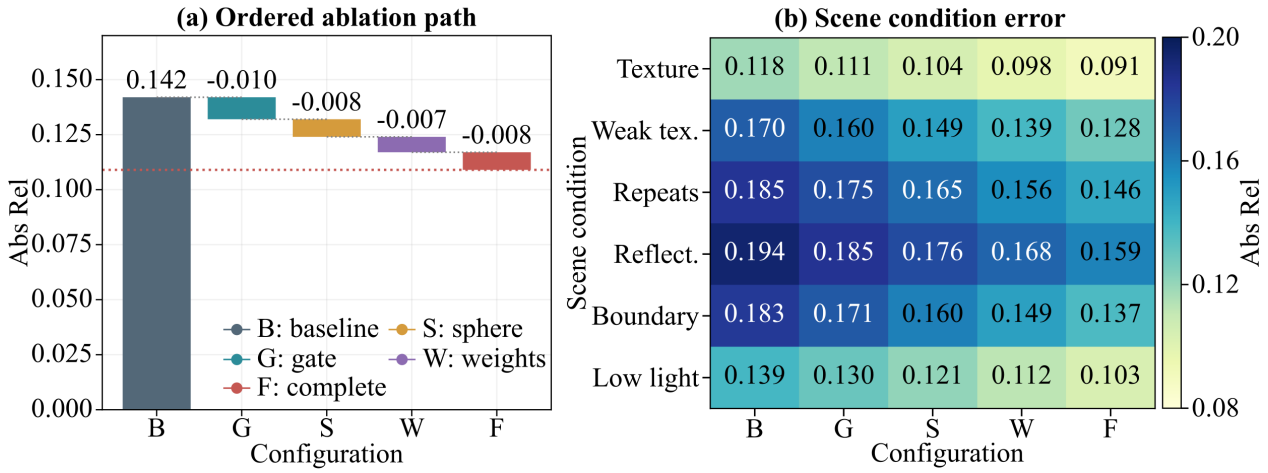


Figure 5. Component and scene condition attribution. (a) Ordered ablation waterfall. (b) Annotated condition by variant error heatmap

Figure 6(a) examines rotation perturbations of 0, 0.1, 0.2, 0.4, 0.6, and 0.8 degrees. Automatic stereo self-calibration methods have estimated varying intrinsic parameters by optimizing correspondence-based geometric costs [27]. The present perturbation test instead varies extrinsic rotation while holding the intrinsic parameters fixed. The complete configuration has errors of 0.109, 0.112, 0.118, 0.133, 0.157, and 0.188. Its error increase accelerates beyond 0.2 degrees, crossing the 0.150 assessment level between 0.4 and 0.6 degrees. This behavior makes support orientation a relevant diagnostic alongside candidate projection, since both operations use the perturbed transformation. Holding the image pair and reference depth fixed isolates the effect of changes in the parameters supplied to candidate projection and support construction.

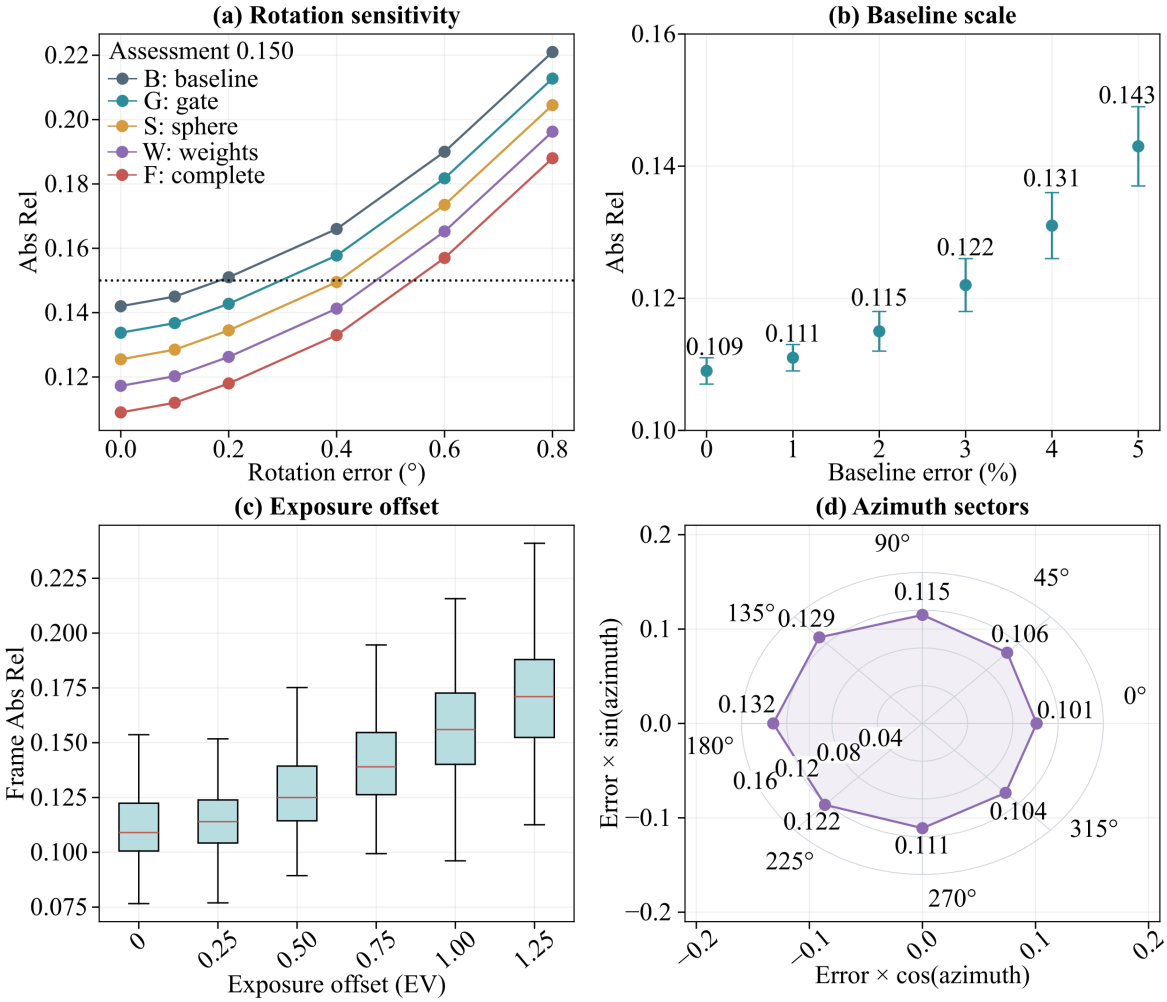


Figure 6. Robustness under calibration and appearance perturbations. (a) Rotation perturbation curves. (b) Baseline scale perturbation scatter with uncertainty bars. (c) Exposure condition boxplots. (d) Azimuth sector polar error profile

Figure 6(b) isolates baseline scale perturbations from 0% to 5%, producing complete model errors of 0.109, 0.111, 0.115, 0.122, 0.131, and 0.143. The 31.2% endpoint increase is smaller than the 72.5% increase across the rotation range, although the two perturbation axes represent different physical changes. The plotted bars are reported as between-run spread, increasing from 0.002 to 0.006. The statistic defining their height and the associated run count are unspecified; interpretation as confidence intervals or significance evidence requires those records. Their widening motivates examination of the relationship between calibration mismatch and variability across optimization runs. Baseline scale perturbation mainly changes the depth to projection relationship, while rotation perturbation changes trajectory direction; separate controls preserve this distinction during interpretation.

Figure 6(c) gives frame error medians of 0.109, 0.114, 0.125, 0.139, 0.156, and 0.171 across exposure offsets from 0 to 1.25 EV. The 56.9% endpoint increase identifies feature stability as a separate sensitivity under exposure perturbation. Continual stereo learning has examined adaptation across heterogeneous weather, road, and city conditions while retaining previously learned scene knowledge [28]. Figure 6(d) presents eight azimuth sector errors spanning 0.101–0.132, a sector range of 0.031. The polar profile describes directional variation in the sector scores. The exposure offsets characterize sensitivity of a fixed model on the supplied scenes; adaptation to changing environments requires a separate evaluation. Interpreting this variation additionally requires sector counts and scene composition, because illumination, surface type, and depth can vary with viewing direction. Extending the evaluation to changing environments would require testing appearance variation alongside camera geometry. Jointly perturbed calibration and exposure would reveal whether these sensitivities reinforce one another on the same frames.

4.3. Computational Sensitivity and Repeatability

Figure 7(a) evaluates 32, 48, 64, 80, 96, and 128 candidates, with errors of 0.153, 0.132, 0.119, 0.112, 0.109, and 0.107. Corresponding latencies are 34, 39, 44, 49, 54, and 67 ms. Efficient stereo networks reduce cost-aggregation demands through adaptive downsampling and disparity alignment, exposing a tradeoff between spatial detail and computational cost [29]. Increasing resolution from 32 to 64 candidates changes error by 0.034 for ten additional milliseconds. Increasing it from 96 to 128 changes error by 0.002 for thirteen additional milliseconds. The latter change gives a substantially smaller marginal return over this candidate range. The selected 96 candidate configuration lies near this transition and preserves a common search grid for the architectural comparisons.

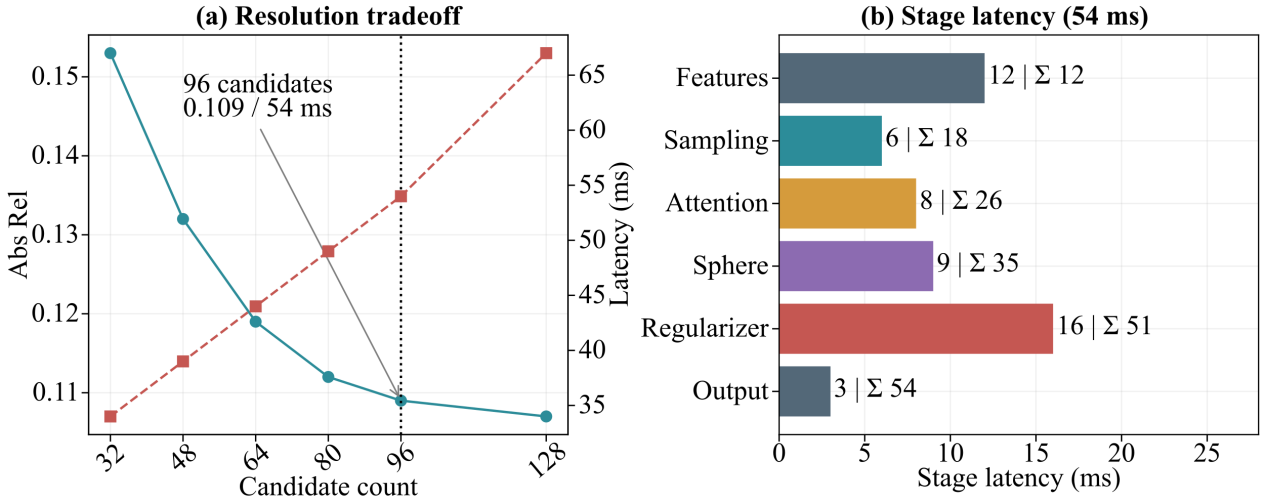


Figure 7. Candidate resolution and runtime accounting. (a) Candidate count trajectories for error and latency. (b) Horizontal stage latency bars with cumulative annotations

Figure 7(b) allocates the 54 ms latency across six stages: 12 ms for feature extraction, 6 ms for coordinate sampling, 8 ms for correlation and attention, 9 ms for spherical aggregation, 16 ms for regularization, and 3 ms for regression and output. Regularization contributes 29.6% of the total, followed by feature extraction at 22.2% and spherical aggregation at 16.7%. Coordinate sampling contributes 11.1%, making irregular access relevant alongside the nominal arithmetic cost of the geometric operations. The stage sum reproduces the complete configuration latency. Implementation profiling should record whether coordinate maps are cached, since caching transfers work from repeated geometry construction to memory storage and feature access.

Reducing regularization time offers a different intervention from reducing candidate resolution. Candidate reduction changes the resolution available throughout the volume, whereas a faster regularizer could preserve the chosen hypothesis grid. At 96 candidates, removing two milliseconds from regularization would reduce total latency by 3.7% under the stage allocation. The same change in coordinate sampling would require a larger fraction of that state's budget. Memory analysis should retain both cached coordinates and temporary sampled tensors, with a common precision and

batch size. Synchronization at stage boundaries and repeated warm executions would connect component timings to the end-to-end latency while accounting for asynchronous execution.

Figure 8(a) presents paired values for five initialization seeds. Baseline values of 0.146, 0.140, 0.143, 0.139, and 0.142 have a mean of 0.142; complete model values of 0.112, 0.107, 0.110, 0.108, and 0.108 have a mean of 0.109. Paired differences range from 0.031 to 0.034, with the same ordering across all five paired comparisons. Their mean is 0.033 and their sample standard deviation is approximately 0.00122 across the five supplied pairs. These descriptive statistics characterize the displayed initialization comparison; dataset-level generalization requires sequence-aware uncertainty analysis. These run-level comparisons complement correspondence-level reliability mechanisms, including confidence-aware adaptive disparity refinement [30]. Seed comparisons retain identical data partitions and test masks while varying initialization and training order under the same optimization schedule.

Figure 8(b) shows unconnected empirical cumulative samples from 120 frame scores for each configuration. Each displayed sample set is scaled to match its configuration-level mean, so the horizontal axis remains absolute relative error rather than a dimensionless unit-mean score. This display constraint separates the graphic from an independently verified empirical distribution of raw frame residuals. A vertical marker at 0.150 and a horizontal marker at cumulative fraction 0.95 identify threshold exceedance and the upper five percent tail. The distributions overlap despite their different means, illustrating why the frame population remains relevant after a favorable aggregate comparison. A particularly difficult frame can retain a large error even when its configuration has the smallest mean. Sequence level resampling would provide a suitable unit for estimating uncertainty when temporally adjacent frames are correlated.

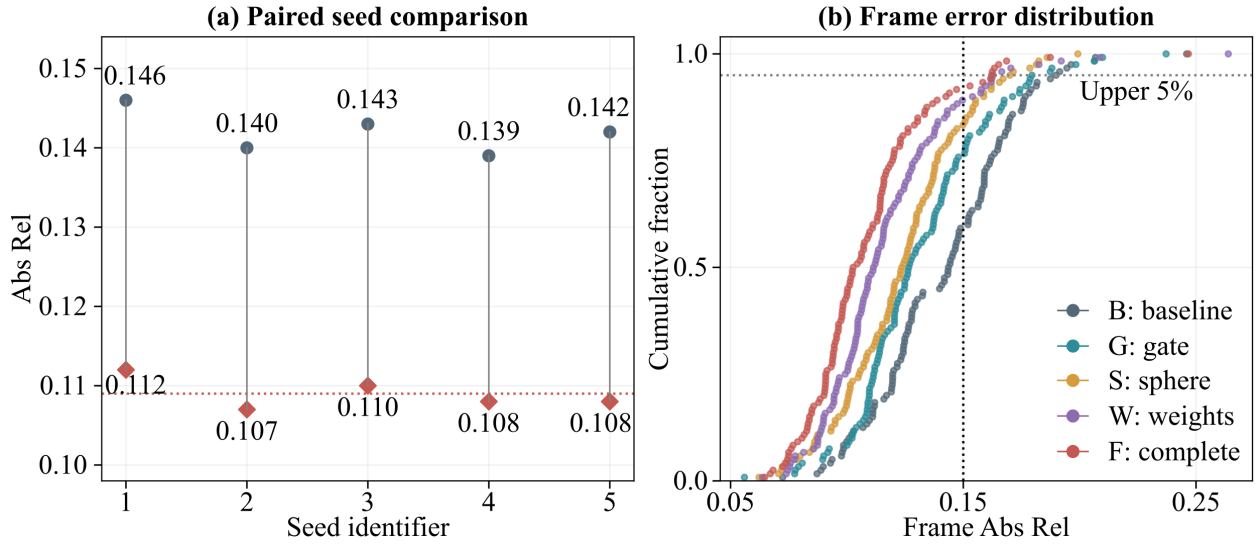


Figure 8. Repeatability and frame error distributions. (a) Paired seed comparisons. (b) Frame error empirical cumulative distributions with threshold and tail annotations

Sequence identity provides a grouping variable for investigating upper tail frame errors. Frames from the same trajectory may share surface appearance, viewing direction, and calibration conditions, making their error patterns dependent on the scene encountered. Confidence intervals should account for this grouping, because resampling individual adjacent frames can overstate the effective number of independent observations. The analysis should pair frames across configurations and record which sequences contribute the largest tail errors. Those frames can then be stratified by incidence angle, depth, visibility, and exposure using the same definitions as the earlier comparisons. Tracking unsupported pixel rate and the duration of high error episodes would further connect frame scores to persistent reconstruction failures along a moving camera trajectory.

5. Conclusion

This study develops a spherical epipolar aggregation extension of ACVNet for calibrated fisheye stereo depth estimation, with the reported numerical comparison suggesting improved matching under geometry-aligned candidate construction and cost propagation.

1. Overall depth estimation accuracy. The complete configuration reduces absolute relative error from 0.142 to 0.109, corresponding to a relative reduction of 23.2%, while root mean square error decreases from 1.84 to 1.43 m. The proportion of pixels exceeding ten percent relative error decreases from 18.6% to 13.4%, indicating improvement in both average accuracy and threshold-based reliability.

2. Peripheral performance and computational tradeoffs. In the outermost incidence angle band, absolute relative error decreases from 0.201 to 0.139, a reduction of 30.8%, compared with 16.5% near the optical axis. This spatial pattern

motivates further evaluation of spherical support alignment at large viewing angles. Potential uses include nearby-obstacle perception on mobile platforms with synchronized, stably calibrated cameras. The accuracy improvement accompanies a latency increase from 46 to 54 ms per stereo pair, exposing a practical tradeoff between reconstruction quality and computational cost.

3. Geometric contribution and operating limitations. The architectural contribution couples calibrated ray candidates with great circle support and validity aware residual fusion. Its effectiveness depends on calibration stability: a rotation perturbation of 0.8 degrees increases absolute relative error from 0.109 to 0.188. Reflective surfaces, thin structures, and multimodal correspondence distributions remain challenging conditions for support weighting and depth regression. Transfer to additional devices and scenes depends on documented dataset provenance, hardware-specific profiling, and independent benchmark validation.

Future work will investigate calibration adaptive support, uncertainty aware prediction, and more efficient candidate allocation to improve robustness within constrained computational budgets.

6. Declarations

6.1 Author Contributions

Leon Kamiński contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Emil Pawlak contributes to validation, analysis, investigation, data collection, draft preparation. Marcin Chmielewski contributes to conceptualization, validation, analysis. Sebastian Zalewski contributes to validation, analysis, investigation. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Date Availability

Not applicable.

6.6 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.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Roghani, S. E. S., & Koyuncu, E. (2024). Canonical trinocular camera setups and fisheye view for enhanced feature-based visual aerial odometry. IEEE Access, 12, 134888–134901. https://doi.org/10.1109/access.2024.3456909

[2] Zhu, H., Zhang, G., Ye, Z., & Zhou, H. (2024). LVIF: A lightweight tightly coupled stereo-inertial SLAM with fisheye camera. Complex & Intelligent Systems, 10(1), 763–780. https://doi.org/10.1007/s40747-023-01190-5

[3] Simioni, E., Pernechele, C., Erb, W., Marchini, A., Martini, P., Lessio, L., Penasa, L., & Beghini, M. (2024). A-central model for the geometric calibration of hyper-hemispherical lenses. Optics Express, 32(20), 34777–34795. https://doi.org/10.1364/oe.527318

[4] Zhou, G., Tang, Y., Xie, Y., Song, R., Zhu, Q., & Zhou, X. (2024). Calibration of fisheye camera with colinear constraint of the main optical axis and epipolar line orthogonality-transverse axis. IEEE Sensors Journal, 24(12), 19373–19384.

https://doi.org/10.1109/jsen.2024.3392810

[5] Zhang, Y., Ma, Y., Wu, Y., & Liu, L. (2024). Achieving widely distributed feature matches using flattened-affine-SIFT algorithm for fisheye images. Optics Express, 32(5), 7969–7986. https://doi.org/10.1364/oe.513531

[6] Jiang, H., Xu, R., Tan, M., & Jiang, W. (2024). RomniStereo: Recurrent omnidirectional stereo matching. IEEE Robotics and Automation Letters, 9(3), 2511–2518. https://doi.org/10.1109/lra.2024.3357315

[7] Chen, Q., Ge, B., & Quan, J. (2024). Unambiguous pyramid cost volumes fusion for stereo matching. IEEE Transactions on Circuits and Systems for Video Technology, 34(10), 9223–9236. https://doi.org/10.1109/tcsvt.2023.3291726

[8] Xu, G., Wang, X., Zhang, Z., Cheng, J., Liao, C., & Yang, X. (2025). IGEV++: Iterative multi-range geometry encoding volumes for stereo matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(8), 7108–7122. https://doi.org/10.1109/tpami.2025.3569218

[9] Xu, G., Wang, Y., Cheng, J., Tang, J., & Yang, X. (2024). Accurate and efficient stereo matching via attention concatenation volume. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4), 2461–2474. https://doi.org/10.1109/tpami.2023.3335480

[10] Liao, Z., Zhou, W., & Li, H. (2024). DaFIR: Distortion-aware representation learning for fisheye image rectification. IEEE Transactions on Circuits and Systems for Video Technology, 34(5), 3606–3618. https://doi.org/10.1109/tcsvt.2023.3315967

[11] Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401–7416. https://doi.org/10.1109/tcsvt.2024.3376692

[12] Chou, C.-H., & Tsao, L.-H. (2025). Automatic calibration of multiple fisheye cameras using recovered human body mesh. Electronic Imaging, 37(10), 347-1–347-8. https://doi.org/10.2352/ei.2025.37.10.ipas-347

[13] Liang, Z., & Li, C. (2024). Any-Stereo: Arbitrary scale disparity estimation for iterative stereo matching. Proceedings of the AAAI Conference on Artificial Intelligence, 38(4), 3333–3341. https://doi.org/10.1609/aaai.v38i4.28119

[14] Wang, S., Li, B., & Dai, Y. (2024). Efficient multi-view stereo by dynamic cost volume and cross-scale propagation. IEEE Transactions on Circuits and Systems for Video Technology, 34(10), 9414–9427. https://doi.org/10.1109/tcsvt.2024.3398060

[15] Zhou, J., Lin, C., Yang, S., Liao, K., & Zhao, Y. (2025). Calibrating the principal point of vehicle-mounted fisheye cameras using point-oriented representation. IEEE Sensors Journal, 25(11), 19602–19612. https://doi.org/10.1109/jsen.2025.3555836

[16] Wang, Y., Wang, L., Li, K., Zhang, Y., Wu, D. O., & Guo, Y. (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

[17] Okae, J., & Qin, H. (2024). DMCNet: Toward lightweight volumetric stereo matching. IEEE Sensors Journal, 24(7), 11936–11945. https://doi.org/10.1109/jsen.2024.3368028

[18] Jin, L., & Xu, K. (2025). MAF-Stereo: Fast stereo matching through multi-branch attention fusion. ISA Transactions, 164, 211–221. https://doi.org/10.1016/j.isatra.2025.05.038

[19] Li, P., Yao, C., Jia, Y., & Wu, Y. (2025). Inter-scale similarity guided cost aggregation for stereo matching. IEEE Transactions on Circuits and Systems for Video Technology, 35(1), 134–147. https://doi.org/10.1109/tcsvt.2024.3453965

[20] Pathak, S., Hamada, T., & Umeda, K. (2024). Trinocular 360-degree stereo for accurate all-round 3D reconstruction considering uncertainty. Advanced Robotics, 38(15), 1038–1051. https://doi.org/10.1080/01691864.2024.2376022

[21] Tosi, F., Bartolomei, L., & Poggi, M. (2025). A survey on deep stereo matching in the twenties. International Journal of Computer Vision, 133(7), 4245–4276. https://doi.org/10.1007/s11263-024-02331-0

[22] Lee, J. Y., Ka, W., Choi, J., & Kim, J. (2024). Modeling stereo-confidence out of the end-to-end stereo-matching network via disparity plane sweep. Proceedings of the AAAI Conference on Artificial Intelligence, 38(4), 2901–2910. https://doi.org/10.1609/aaai.v38i4.28071

[23] Xu, Y., & Dong, Q. (2026). Distortion-aware depth self-updating for self-supervised fisheye monocular depth estimation. IEEE Transactions on Image Processing, 35, 1883–1898. https://doi.org/10.1109/tip.2026.3661813

[24] Zhang, C., Meng, G., Tian, K., Ni, B., & Xiang, S. (2024). Active disparity sampling for stereo matching with adjoint network. IEEE Transactions on Image Processing, 33, 354–365. https://doi.org/10.1109/tip.2023.3343105

[25] Zhu, L., Rigall, E., Gao, Y., Zhang, Z., Bai, Y., & Dong, J. (2025). Region-aware driven distribution optimization for stereo matching. IEEE Transactions on Circuits and Systems for Video Technology, 35(12), 11965–11977. https://doi.org/10.1109/tcsvt.2025.3588869

[26] Su, W., & Tao, W. (2025). Context-aware multi-view stereo network for efficient edge-preserving depth estimation. International Journal of Computer Vision, 133(6), 3367–3391. https://doi.org/10.1007/s11263-024-02337-8

[27] Satouri, B., El Abderrahmani, A., & Satori, K. (2024). End-to-end artificial intelligence-based system for automatic stereo camera self-calibration. IEEE Access, 12, 160927–160945. https://doi.org/10.1109/access.2024.3486017

[28] Zhang, C., Meng, G., Fan, B., Tian, K., Zhang, Z., Xiang, S., & Pan, C. (2024). Reusable architecture growth for continual stereo matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(9), 6167–6184. https://doi.org/10.1109/tpami.2024.3378884

[29] Wang, Y., Li, K., Wang, L., Hu, J., Wu, D. O., & Guo, Y. (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

[30] Park, C., Kim, J., Kweon, M., & Park, J. (2026). CAR-Stereo: Confidence-aware adaptive disparity refinement for real-time stereo matching. IEEE Robotics and Automation Letters, 11(1), 690–697. https://doi.org/10.1109/lra.2025.3634908