agriculture

MDPI

Article

MI-ACVNet: A Lightweight Stereo Matching Network for High-Precision Single-View 3D Reconstruction of Kirin Watermelons

Zetong Li [1,2,3,4], Xufeng Xu [1,2,3,4], Yuan Gao [1,2,3,4], Wenqian Lei [1,2,3,4] and Xiuqin Rao [1,2,3,4,*]

[1] College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China; 22313051@zju.edu.cn (Z.L.); 12213062@zju.edu.cn (X.X.); yuang@zju.edu.cn (Y.G.); 22513045@zju.edu.cn (W.L.)
[2] State Key Laboratory of Agricultural Equipment, Hangzhou 310058, China
[3] Key Laboratory of on Site Processing Equipment for Agricultural Products, Ministry of Agriculture, Hangzhou 310058, China
[4] Key Laboratory of Intelligent Equipment and Robotics for Agriculture of Zhejiang Province, Hangzhou 310058, China
* Correspondence: xqrao@zju.edu.cn

Abstract

Three-dimensional surface reconstruction is essential for accurately acquiring the external quality parameters of watermelons, such as size, volume, and defect area. Binocular stereo vision provides a low-cost and easily deployable solution for the single-view 3D reconstruction of watermelons. However, watermelons present highly similar surface textures, and as typical spheroid-like objects, the excessive angle between surface normals of edge regions and the camera optical axis leads to insufficient feature representation. Consequently, directly applying existing stereo matching algorithms often introduces matching ambiguities, and lightweight networks struggle to balance real-time performance with matching accuracy. This study focuses on the high-precision single-view point cloud generation of Kirin watermelons. To address these issues, we first construct a cross-modal, high-precision Kirin watermelon stereo matching dataset. Building upon the Fast-ACVNet+ architecture, we then propose MI-ACVNet, a lightweight stereo matching network tailored for high-precision watermelon point cloud acquisition. In the feature extraction stage, a Multi-Scale Stereo Feature Extraction (MSFE) module is adapted. By incorporating the re-parameterized network MobileOne and Epipolar-Enhanced Coordinate Attention (E2CA), MSFE improves the discriminative capability for weak and similar textures without compromising inference speed. For cost computation, a Coarse-to-Fine Cascaded Residual Correction (C2F-CRC) strategy is incorporated to construct a fine-grained cost volume via sub-pixel interpolation, enhancing the network's ability to capture subtle surface fluctuations. Furthermore, a Semantics-Guided Region-Aware Loss (SGRA-Loss) is formulated, leveraging semantic masks to apply differentiated supervision weights across edge, center, and background regions to significantly improve edge matching accuracy. Ablation studies validate the effectiveness of the MSFE, C2F-CRC, and SGRA-Loss components. Compared to the baseline model, the full MI-ACVNet reduces the End-Point Error (EPE) by 19.5% and the Bad-0.5 error rate by 34.5% in the watermelon region. Furthermore, when compared against five mainstream algorithms (StereoNet, AANet, HSMNet, LightStereo-L, and NMRF-swint), MI-ACVNet achieves state-of-the-art performance: EPE and Bad-0.5 are reduced to 0.091 pixels and 1.159%, respectively, with a single-frame inference time of only 46 ms. The average depth error of the reconstructed point clouds is merely 0.26 mm. By ensuring both real-time efficiency and high-precision depth estimation, this method demonstrates promising potential for deployment in industrial Kirin watermelon sorting lines, driving sorting equipment toward higher precision and intelligence.

Academic Editors: Dong-Hoon Lee and Yongjin Cho

Received: 9 June 2026
Revised: 1 July 2026
Accepted: 3 July 2026
Published: 6 July 2026

Keywords: watermelon; deep learning; stereo matching; 3D reconstruction; non-destructive testing

1. Introduction

Watermelons are a vital economic crop that holds a significant position in the global fruit and vegetable market [1]. With the continuous improvement in living standards, consumer demand for high-quality, standardized branded watermelons is growing, imposing higher requirements for post-harvest commercial processing [2]. Parameters such as fruit shape, size, volume, and surface defect area are crucial grading indicators [3]. Traditionally, watermelon grading has relied heavily on manual sorting, which is inefficient, highly subjective, and inadequate for modern large-scale agricultural production [4]. In recent years, with the rapid development of artificial intelligence, machine vision has become a core technology in modern agricultural sensing systems. Advanced vision algorithms have been widely applied in agricultural robots for navigation in greenhouse and field environments [5] and in image-based high-throughput plant phenotyping [6]. These technologies are transforming many aspects of agricultural production and management. In the postharvest processing of agricultural products, machine vision-based nondestructive inspection has also gradually replaced manual sorting [7]. Early machine vision systems primarily relied on 2D image analysis, estimating volume and weight by extracting projection features like area and diameter [8]. However, this dimensionality reduction inherently discards depth information and surface geometry during projection, severely limiting its accuracy in size measurement, volume estimation, and defect area quantification.

To overcome the limitations of 2D vision, 3D surface reconstruction technology has emerged. Constructing digital 3D models allows for comprehensive acquisition of geometric and textural data, enabling precise measurement of diameter, volume, density, and defect area [9] to reliably support quality assessment. In recent years, three mainstream technical routes have emerged in the field of fruit nondestructive testing according to the principle of depth acquisition: depth camera-based methods [10–12], image sequence-based methods [13–15], and binocular stereo vision-based methods [16–18]. Depth camera-based approaches (e.g., ToF cameras [19], consumer-grade speckle structured light [20]) feature high integration but are limited by sensor resolution and systematic errors [21], with depth measurement accuracy typically at the millimeter level. Industrial-grade coded structured light delivers ultra-high precision yet suffers from high equipment cost and long acquisition time, failing to meet the cost and speed requirements of online sorting. Image sequence-based methods (e.g., SfM-MVS [22], NeRF [23], 3DGS [24]) exhibit remarkable geometric accuracy. In particular, recent neural 3D representation techniques, such as Neural Radiance Fields (NeRF), have shown great potential for high-fidelity 3D reconstruction and novel view synthesis in complex scenes [25]. However, for online fruit sorting, these methods rely heavily on the acquisition of a large number of multi-view images. They also require complex feature matching or iterative optimization, resulting in long reconstruction times. Therefore, they cannot meet the high-throughput and low-latency requirements of fruit sorting [26]. In contrast, binocular stereo vision mimics human binocular mechanisms by computing disparity between stereo pairs to recover 3D shapes. It offers low hardware costs, high potential accuracy, and robust real-time performance [18]. Combined with high-resolution cameras and advanced stereo matching algorithms, this method can deliver millisecond-level, high-precision surface reconstruction, making it highly promising for fruit sorting equipment.

As the core of a binocular vision system, stereo matching algorithms directly determine the accuracy and efficiency of reconstruction. Traditional stereo matching methods (e.g., SGBM [27]) rely heavily on local texture gradient information, and are prone to matching ambiguity and large-scale disparity holes when applied to scenes such as watermelon peels with large smooth regions or repetitive stripes. In recent years, end-to-end deep learning networks (e.g., PSMNet [28], ACVNet [29], MoCha-Stereo [30]) and related deep learning-based image restoration methods [31] have significantly improved dense visual estimation and low-level image analysis accuracy. However, such models typically involve extensive 3D convolution computations or iterative optimization strategies, making it difficult to meet real-time requirements. To balance speed and precision, a series of lightweight stereo matching networks (e.g., StereoNet [32], Fast-ACVNet [33], LightStereo [34]) have been proposed successively. However, directly applying existing lightweight networks to watermelon surface reconstruction still faces two critical problems. First, weak and similar textures cause interference: lightweight networks typically adopt downsampling for faster inference, which tends to lose fine texture details on watermelon peels. Second, edge matching accuracy is insufficient: watermelons are spheroid-like objects, and the large angle between the surface normals of edge regions and the camera optical axis sharply reduces the number of effective pixels per unit physical area. This leads to inadequate feature representation at the periphery. Existing lightweight networks struggle to maintain real-time performance while ensuring matching accuracy in such scenarios.

To address these issues, a lightweight stereo matching network based on cascaded residual correction, named MI-ACVNet (Mobile Incremental Attention Cost Volume Network), is proposed for the high-precision single-view 3D surface reconstruction of Kirin watermelons. By introducing improved strategies including Epipolar-Enhanced Multi-Scale Stereo Feature Extraction, Coarse-to-Fine Cascaded Residual Correction, and Semantics-Guided Region-Aware Loss, the network effectively enhances matching accuracy in weak-texture and edge regions. This work can provide algorithmic support for high-precision intelligent watermelon sorting equipment, promoting the advancement and practical deployment of postharvest commercial processing for fruits and vegetables.

2. Materials and Methods

2.1. Experimental Materials

A total of 101 Kirin watermelon samples, exhibiting significant variations in size, shape, and surface stripe distribution, were procured from a fruit market in Hangzhou, Zhejiang Province, in March 2025. Prior to the experiments, all samples were equilibrated at room temperature for 24 h.

2.2. Image Acquisition System

A multi-view binocular image acquisition system was established. The overall structure of the system is shown in Figure 1. It mainly consists of five groups of binocular cameras (ZED Mini, Stereolabs, San Francisco, California, USA; maximum binocular resolution 4416 × 1242 pixels), one group of high-precision industrial 3D structured-light camera (Mech-Eye PRO S, Mech-Mind Robotics, Beijing, China; 1920 × 1200 pixels, accuracy: 0.05 mm @ 1 m), and a cylindrical uniform light box integrated with light homogenizing film, reflective paper and LED light sources. The five groups of binocular cameras adopt an orthogonal layout. One group is placed vertically above the watermelon for downward shooting, and arranged compactly in parallel with the structured-light camera. The other four groups are orthogonally distributed at 90° intervals around the horizontal plane to capture surface information from multiple perspectives. Although the system is designed for full-surface 3D reconstruction, this study focuses specifically on high-precision point

cloud generation from a single viewpoint. Consequently, only the top core acquisition unit (i.e., the top-view binocular camera and the structured light camera) was utilized for subsequent dataset construction and model validation.

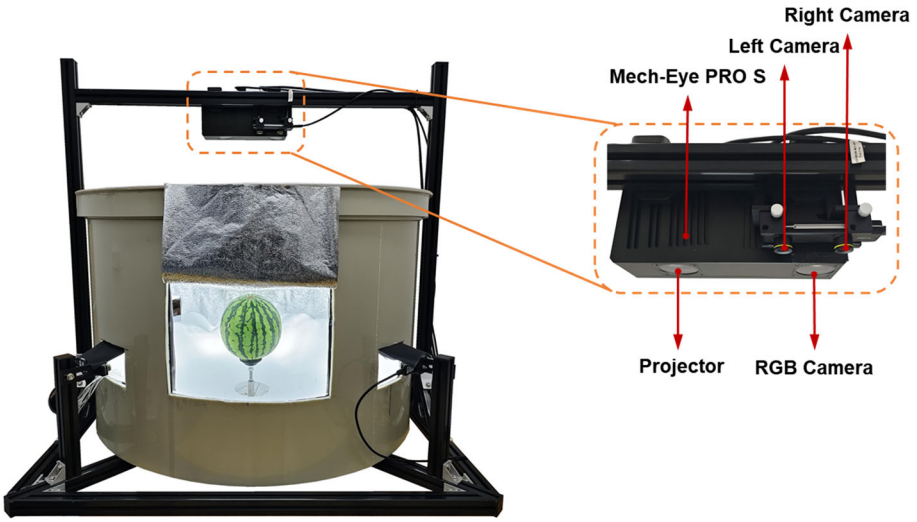


Figure 1. Multi-view binocular image acquisition system for watermelons.

The training and validation of the stereo matching network were conducted on a high-performance deep learning computing server. The hardware configuration includes dual Intel Xeon Platinum 8368 processors, 256 GB of RAM, and eight NVIDIA GeForce RTX 4090 GPUs, providing ample computational power for the training and efficient inference of complex deep learning models. The core algorithm was developed using Python 3.9 within the PyTorch (version 1.12.1) deep learning framework. PyCharm (version 2023.1.1) was utilized as the integrated development environment (IDE), and the system environment was managed via Anaconda (version 24.1.2).

2.3. Construction of the Kirin Watermelon Stereo Matching Dataset

A standard stereo matching dataset comprises left and right views along with their corresponding disparity maps. In this study, the left and right RGB views were acquired using the ZED Mini binocular camera, while the disparity maps were converted from depth maps based on binocular vision principles. Although the ZED Mini camera features a built-in depth estimation function, its integrated algorithm possesses limited accuracy and cannot serve directly as ground-truth labels for training. Therefore, the Mech-Mind Mech-Eye PRO S industrial 3D camera was selected to acquire the ground truth. This camera integrates a high-intensity structured light projector and a high-resolution RGB camera, achieving a depth measurement accuracy of 0.05 mm at 1 m. As illustrated in Figure 1, the spatial relative position of the two camera sets adopts a compact side-by-side layout with parallel optical axes. This setup ensures a sufficient common field of view between the depth camera and the binocular camera.

2.3.1. Dataset Construction Pipeline

The core of dataset construction is to accurately map depth information from the structured-light camera coordinate system to the pixel coordinate system of the binocular left camera and convert it into a disparity map. The specific pipeline is illustrated in Figure 2, which comprises three key steps:

(1) oint Calibration: A high-precision aluminum checkerboard calibration plate (specification: 12 × 9 corners, square side length 20 mm) is employed as the spatial reference.

Image pairs of the calibration plate are captured separately by the ZED Mini left camera and the Mech-Eye structured-light camera. Based on the PnP (Perspective-n-Point) algorithm, the extrinsic matrix between the structured-light camera coordinate system O_{mech} and the binocular left camera coordinate system O_{ZED} is solved, i.e., the rotation matrix $R_{mech \to ZED}$ and translation vector $t_{mech \to ZED}$.

(2) Spatial Reprojection: First, utilizing the intrinsic matrix and the raw depth map of the structured light camera, pixel coordinates are back-projected to reconstruct a 3D point cloud $P_S = (X_S, Y_S, Z_S)^T$ in its camera coordinate system. Subsequently, utilizing the extrinsic matrix derived from the joint calibration, the point cloud P_S is transformed into the left camera coordinate system to obtain $P_L = (X_L, Y_L, Z_L)^T$.

$$P_L = R_{mech \to ZED} \cdot P_S + T_{mech \to ZED} \tag{1}$$

Next, based on the intrinsic matrix K_{ZED} of the ZED Mini left camera, the transformed point cloud P_L is forward-projected onto the left image plane (u, v). The Z_L component of the point cloud is extracted as the depth value for the corresponding pixel, thereby generating a high-precision depth map strictly aligned with the left stereo view.

(3) Disparity Conversion: Based on binocular imaging principles, the left depth map generated in the previous step was converted into a disparity map d using the baseline B of the ZED Mini camera and the focal length f of the left camera:

$$d = \frac{f \cdot B}{Z} \tag{2}$$

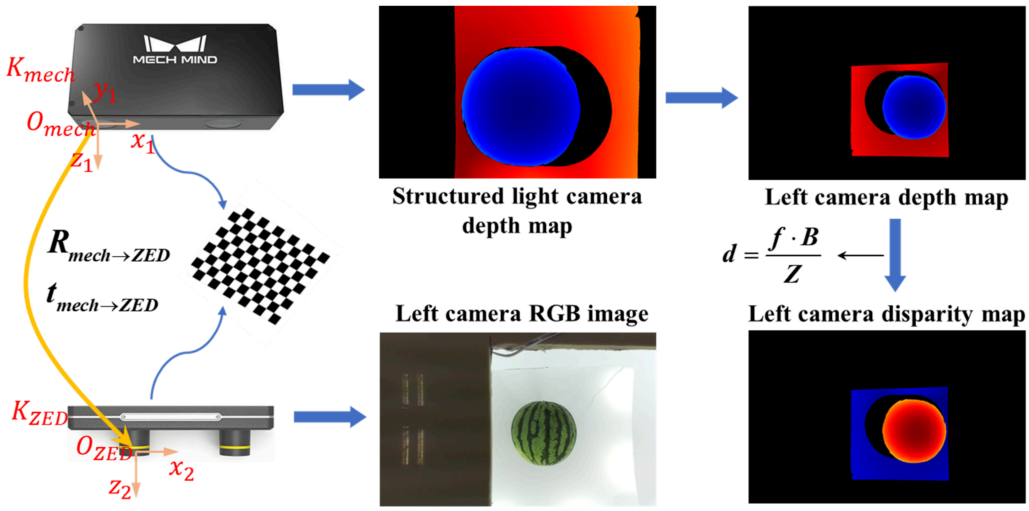


Figure 2. Flowchart of high-precision disparity map generation. In the depth and disparity maps, blue and red colors represent smaller and larger values, respectively.

Through the aforementioned processing, dense disparity ground truths pixel-aligned with the ZED Mini left view were ultimately obtained. A watermelon stereo matching dataset comprising 808 pairs of stereo images and their corresponding disparity ground truths was constructed.

2.3.2. Dataset Accuracy Validation and Evaluation

Due to viewpoint discrepancies between the binocular and structured light cameras, coupled with inherent stereo calibration errors, rigorous accuracy validation of the generated dataset is imperative to ensure ground truth reliability. Consequently, 20 pairs of

calibration board images in varying poses were captured for validation, with the sampling range covering both the center and the four corners of the field of view (FOV). The validation scheme encompasses two dimensions: XY-plane reprojection error and Z-axis depth geometric error.

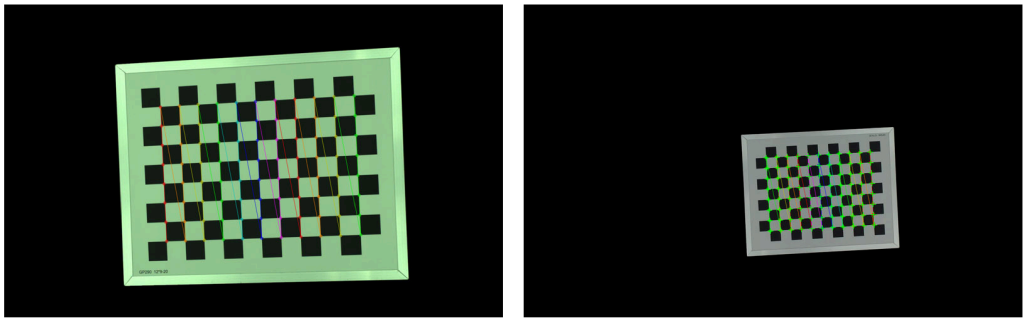
1. XY-Plane Reprojection Error Validation

To evaluate the alignment accuracy between the reprojected left depth map and the original left RGB image, a reprojection test was conducted utilizing the RGB texture information from the structured light camera. The specific steps are as follows:

(1) Corner Extraction: First, extract the corner pixel coordinates of the calibration board, denoted as p_{mech}, from the RGB image of the structured light camera.
(2) Coordinate Mapping: Utilizing the reprojection method detailed in the previous section, map and project p_{mech} onto the pixel plane of the left camera to derive the theoretical projection coordinates p_{proj}.
(3) Error Calculation: Extract the corresponding actual corner coordinates p_{ZED}, from the captured left RGB image. The Euclidean distance between the theoretical and actual coordinates is calculated as the planar reprojection error:

$$E_{xy} = \|p_{proj} - p_{ZED}\|_2 = \sqrt{(u_{proj} - u_{ZED})^2 + (v_{proj} - v_{ZED})^2} \quad (3)$$

Statistical analysis of the 20 calibration board images reveals an average reprojection error of merely 0.56 pixels. To visually demonstrate the alignment quality post-reprojection, a representative sample is shown in Figure 3. Figure 3a presents the corner extraction results from the structured light camera's perspective, whereas Figure 3b illustrates the projection of these extracted corners onto the left stereo view. The green crosses, representing the reprojected corner positions, precisely coincide with the checkerboard corners in the original image without any visually perceptible deviation. This strongly verifies that the depth map transformed from the structured light camera achieves sub-pixel alignment with the left RGB image.



(**a**) (**b**)

Figure 3. Visualization of planar reprojection results: (**a**) Corner extraction from the structured light camera source image; (**b**) Reprojection effect on the ZED left view.

2. Z-axis Depth Geometric Error Verification

The accuracy of the depth map directly determines the quality of the disparity ground truth. To verify the precision of the generated left-view depth map, a calibration method leveraging the inherent geometric dimension constraints of the calibration plate is proposed. The specific steps are as follows:

(1) Spatial Distance Calculation: In the generated left-view depth map, select the pair of most distant corner points from the calibration plate corner array (i.e., the top-left and bottom-right corners). Using the depth values and camera intrinsic parameters,

compute their 3D spatial coordinates in the left camera coordinate system, and calculate the Euclidean distance D_{calc}.

(2) Ground Truth Comparison: is compared against the actual physical distance, D_{gt} (derived from the 20 mm checkerboard grid size), to calculate the relative dimensional error:

$$E_z = \frac{|D_{calc} - D_{gt}|}{D_{gt}} \times 100\% \quad (4)$$

Statistical results show that the average relative error of the generated depth map in geometric dimension measurement is only 0.08%. This result confirms that the left-view depth map and the converted disparity map exhibit extremely high accuracy, fully satisfying the training requirements of the stereo matching network.

It should be noted that the above validation was conducted on a planar calibration board, whereas the actual annotation targets are curved watermelon surfaces. It is therefore necessary to clarify whether these accuracy metrics remain applicable. The extrinsic matrix (R, t) describes a rigid-body transformation between the two camera coordinate systems. Once calibrated, this transformation holds identically for all spatial points regardless of the scene geometry, planar or curved. Furthermore, the depth measurements originate from the Mech-Mind Mech-Eye PRO S structured-light camera, whose factory-rated accuracy of 0.05 mm at 1 m is an intrinsic property of the active structured-light sensing principle and does not rely on surface texture. Consequently, the sub-pixel reprojection accuracy (0.56 px) and the depth geometric accuracy (0.08%) verified on the calibration board are equally valid for watermelon surfaces. These metrics thus serve as reliable evidence for the quality of the generated disparity ground truth.

2.4. Lightweight Stereo Matching Network MI-ACVNet for 3D Surface Reconstruction of Watermelons

2.4.1. Architecture of MI-ACVNet

After investigating the performance and architectures of mainstream stereo matching algorithms, this study selects Fast-ACVNet+ as the baseline model, which balances accuracy and real-time performance. However, although the original network effectively reduces computational redundancy via the attention mechanism, its discrete Top-k-based disparity selection strategy complicates the subsequent disparity prediction process. To address this limitation, a lightweight stereo matching network MI-ACVNet (Mobile Incremental Attention Cost Volume Network) based on cascaded residual correction is proposed. This network retains the efficient attention branch of the baseline while reconstructing its disparity output module. As illustrated in Figure 4, MI-ACVNet inherits the coarse-to-fine design paradigm, with key improvements in feature encoding and disparity regression mechanisms. The overall architecture consists of four main stages:

(1) Multi-Scale Feature Extraction: A Multi-Scale Stereo Feature Extraction (MSFE) module, based on epipolar enhancement, is adopted to encode the left and right views. The re-parameterized network MobileOne [35] replaces the baseline's MobileNetV2 [36]. Combined with epipolar-enhanced attention, a multi-scale feature pyramid is constructed to strengthen the representation of watermelon textures.

(2) Initial Disparity Estimation (Attention branch): Following the paradigm of Fast-ACVNet+, low-resolution features are utilized to construct a correlation volume and aggregate global information. Distinctively, instead of generating sparse indices, MI-ACVNet directly regresses an initial disparity map to serve as a geometric prior for subsequent disparity optimization.

(3) Residual Correction (Main Branch): Guided by the initial disparity, a local search window is established. By introducing a local sub-pixel shift mechanism, a continuous

high-resolution cost volume is constructed via warping operations and bilinear interpolation, replacing the original discrete concatenated volume to predict the disparity residual map.

(4) Output and Supervision: The final disparity is obtained by summing the initial disparity and the residual map. During training, a Semantics-Guided Region-Aware Loss (SGRA-Loss) is introduced, utilizing semantic masks to impose focused supervision on the edge regions of the watermelons.

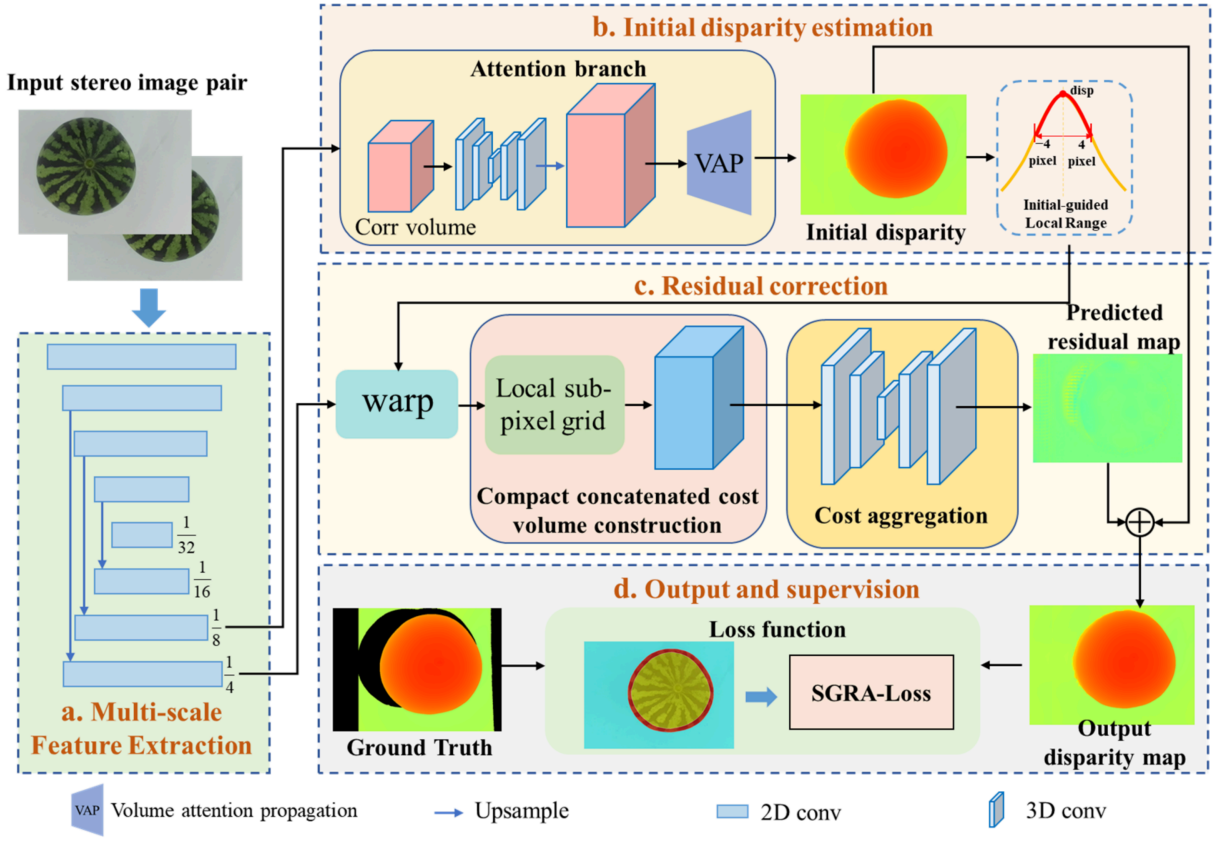


Figure 4. Schematic diagram of the MI-ACVNet architecture.

2.4.2. Optimization of the Feature Extraction Module

The original Fast-ACVNet+ model employs MobileNetV2 for feature extraction and fuses multi-scale features via simple concatenation and convolution. However, this generic design has two key limitations. First, the inverted residual structure of MobileNetV2 tends to lose high-frequency stripe details on watermelon surfaces during downsampling. Second, traditional feature fusion ignores the geometric characteristics of stereo matching: isotropic convolution kernels fail to effectively exploit contextual information along horizontal epipolar lines, leading to matching ambiguity in repetitive texture regions. To address these issues, a Multi-Scale Stereo Feature Extraction module (MSFE) with epipolar enhancement is integrated. As illustrated in Figure 5, this module consists of two main components: a reparameterized backbone network and an Epipolar-Enhanced Coordinate Attention (E2CA) mechanism.

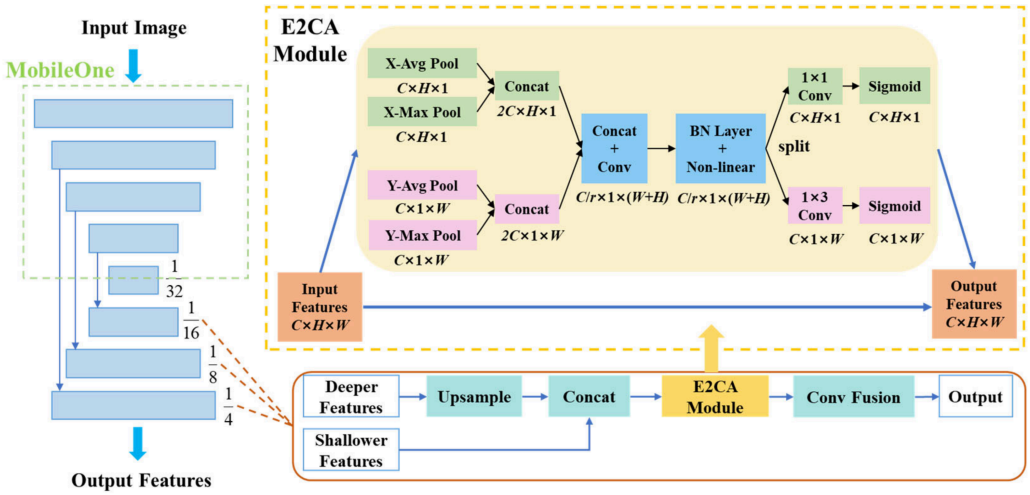


Figure 5. Schematic diagram of the MSFE.

1. Re-parameterized Backbone Network

To enhance feature representation without sacrificing inference speed, this study adopts the MobileOne backbone to replace the original MobileNetV2. The core advantage of MobileOne lies in its decoupled reparameterization mechanism for training and inference: during the training phase, multiple parallel branches are utilized to enrich the gradient flow, strengthening the nonlinear fitting ability for complex textures; during inference, the linear additivity of convolution is leveraged to fuse multi-branch structures into a single convolutional layer without performance loss. This design enables the model to maintain ultra-low latency while significantly improving its capability to characterize edge and texture features.

2. Epipolar-Enhanced Coordinate Attention (E2CA)

To address semantic misalignment and the lack of geometric constraints during feature fusion, this study introduces an enhanced coordinate attention mechanism termed Epipolar-Enhanced Coordinate Attention (E2CA), which spatially reweights the concatenated features, as shown in Figure 5. Compared with the standard Coordinate Attention (CA) module [37], E2CA incorporates two targeted improvements specifically tailored to the stereo matching task: First, to mitigate the blurring of texture edges caused by average pooling, a MaxPool branch is introduced. By fusing average and max features, the network can capture both global context and local high-frequency edges, enhancing the response to watermelon stripe boundaries. Second, according to the principle of stereo vision, disparity search is strictly performed along horizontal epipolar lines. Therefore, in the horizontal encoding branch of E2CA, a 1×3 strip convolution is adopted to replace the original 1×1 convolution (as illustrated in Figure 5). This expands the receptive field along the epipolar direction, strengthens the uniqueness of features in the horizontal direction, while the vertical branch retains a lightweight design.

2.4.3. Coarse-to-Fine Cascaded Residual Correction Strategy (C2F-CRC)

The attention branch in the original Fast-ACVNet+ generates a disparity probability distribution from initial computations and adopts a Top-k selection strategy, in which only the k discrete disparity indices with the highest probabilities are chosen from the full disparity range to construct a compact cost volume. Although this mechanism drastically reduces the search space, it essentially acts as a discrete point-sampling scheme that forcibly discards all spatial information beyond the k isolated points. It also ignores the continuous distribution nature of disparity in real physical space, making it difficult for the model

to capture sub-pixel fluctuations on smooth object surfaces. Inspired by the iterative optimization philosophy widely adopted in stereo matching [38], this study abandons the original Top-k selection strategy and proposes a single-pass cascaded residual correction strategy. Unlike iterative methods that perform multiple refinement loops with shared weights, the proposed C2F-CRC employs a two-stage cascaded architecture. It decomposes the disparity prediction process into two phases: initial estimation and residual correction. Meanwhile, sub-pixel interpolation is introduced to construct a fine-grained cost volume, enabling high-precision matching in weak-texture regions on watermelon surfaces.

1. Cascaded Residual Regression Architecture

To fully exploit the global contextual information extracted by the attention branch, this study reconstructs the prediction flow structure of the network, as shown in Figure 6.

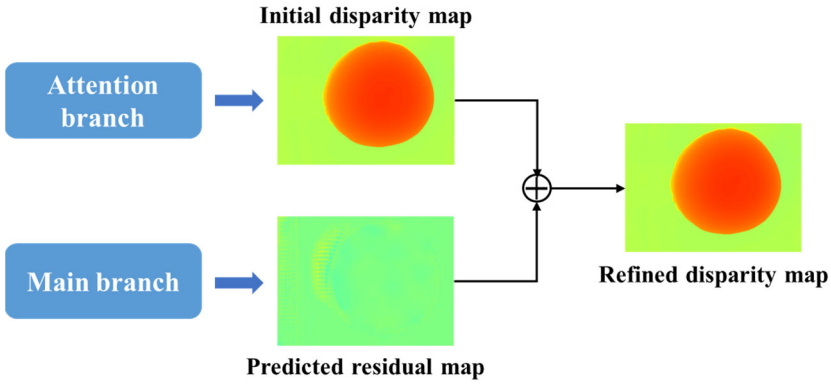


Figure 6. Schematic diagram of the cascaded residual regression architecture.

The reconstructed network consists of two cascaded branches:

(1) Initial Disparity Prediction: The attention branch directly regresses the initial disparity map d_{init}, via a Soft-argmax operation. Although structurally accurate at a macroscopic level, this disparity map lacks fine details in edge and weak-texture regions.

(2) Residual Correction: Instead of directly predicting the final disparity, the main branch takes the initial disparity d_{init} as a prior guide, searches for the optimal offset within a local neighborhood, and focuses on learning the disparity residual Δd. Finally, the high-precision disparity map is obtained by pixel-wise summation of the initial disparity and the predicted residual:

$$d_{final} = d_{init} + \Delta d \tag{5}$$

This residual learning mechanism effectively reduces the optimization difficulty of directly regressing large-range disparities, enabling the model to concentrate on fine-grained detail correction.

2. Fine-Grained Cost Volume Construction Based on Sub-pixel Interpolation

The original Top-k strategy in Fast-ACVNet+ selects discrete disparity candidates based on probability ranking. This filtering mechanism results in a cost volume that is unordered and discontinuous along the disparity dimension, significantly increasing the difficulty for the model to learn the disparity regression distribution. To address this issue, an initial disparity-guided mechanism is first introduced: a fixed local search window $[d_{init} - 4, d_{init} + 4]$ is defined centered on the initial disparity d_{init} (u, v), forcing the cost volume to be constructed in continuous physical space. However, if only integer sampling is performed within this range, the number of feature channels in the cost volume is drastically reduced to 9 (i.e., ± 4 intervals plus the center), leading to a severe shortage of

effective feature information. The network struggles to extract texture details sufficient for high-precision matching from such sparse features. To tackle this limitation, a fine-grained cost volume construction method based on sub-pixel interpolation is employed. The disparity sampling interval is set to 0.25 pixels, and the Bilinear Interpolation algorithm is employed to compute the precise feature response at each sub-pixel position by weighting the feature values of the four neighboring pixels around the sampling point, expanding the number of feature channels in the cost volume from 9 to 33.

2.4.4. Semantics-Guided Region-Aware Weighted Loss Function

In stereo matching tasks, traditional pixel-level loss functions, such as L1 Loss, typically assign equal weights to all pixels. However, for the current task, the matching difficulty and importance vary significantly across different image regions. The central region of the watermelon features relatively smooth textures, making it easy to match. Conversely, the edge region presents heightened matching difficulty due to abrupt disparity variations and occlusions caused by the spherical curvature. The background region, meanwhile, contributes little to the 3D reconstruction. If an equal-weight loss function is employed, the network inherently tends to fit the large flat areas while neglecting the difficult-to-optimize edge details. While spatially weighted loss functions have been explored in prior stereo matching work [18], existing methods typically apply differential weights only between foreground and background regions. This is insufficient to address the problem of large matching errors at object edges. To overcome this limitation, this section extends the differential weighting paradigm into the object interior. A finer-grained region decomposition is introduced, partitioning the image into three regions: object center, object edge, and background.

To precisely partition the image regions, the YOLO11s-seg [39] instance segmentation model is introduced to extract the semantic masks of the watermelons. The specific partitioning pipeline is illustrated in Figure 7. First, the left view is input into the YOLO11s-seg model to obtain the full watermelon region mask M_{full}. Subsequently, morphological erosion is applied to M_{full} to derive the center region mask M_{center}. The annular edge region mask M_{edge} is then generated by subtracting M_{center} from M_{full}. The remaining area outside M_{full} is designated as the background region M_{bg}. Based on this regional division, a pixel-wise weight map $W(u, v)$ is constructed. To compel the network to prioritize the challenging edge details, the highest penalty weight is assigned to the edge region, followed by the center region, with the lowest weight assigned to the background region. The weight allocation strategy is formulated as follows:

$$W(u,v) = \begin{cases} \lambda_{edge}, & \text{if } (u,v) \in M_{edge} \\ \lambda_{center}, & \text{if } (u,v) \in M_{center} \\ \lambda_{bg}, & \text{otherwise} \end{cases} \tag{6}$$

where $\lambda_{edge} > \lambda_{center} > \lambda_{bg}$. The specific values will be determined in subsequent experiments. The MI-ACVNet incorporates two output nodes, corresponding to the initial disparity d_{init} from the attention branch and the disparity residual Δd from the main branch, respectively. To ensure accurate geometric feature learning at each stage, a joint supervision strategy is employed. The attention branch is responsible for outputting an initial disparity with a roughly correct global structure. The loss function for this component computes

the weighted difference between the predicted initial disparity d_{init} and the ground-truth disparity d_{gt}:

$$L_{init} = \frac{1}{N} \sum_{(u,v)} W(u,v) \cdot \text{Smooth}_{L1}\left(d_{init}(u,v) - d_{gt}(u,v)\right) \tag{7}$$

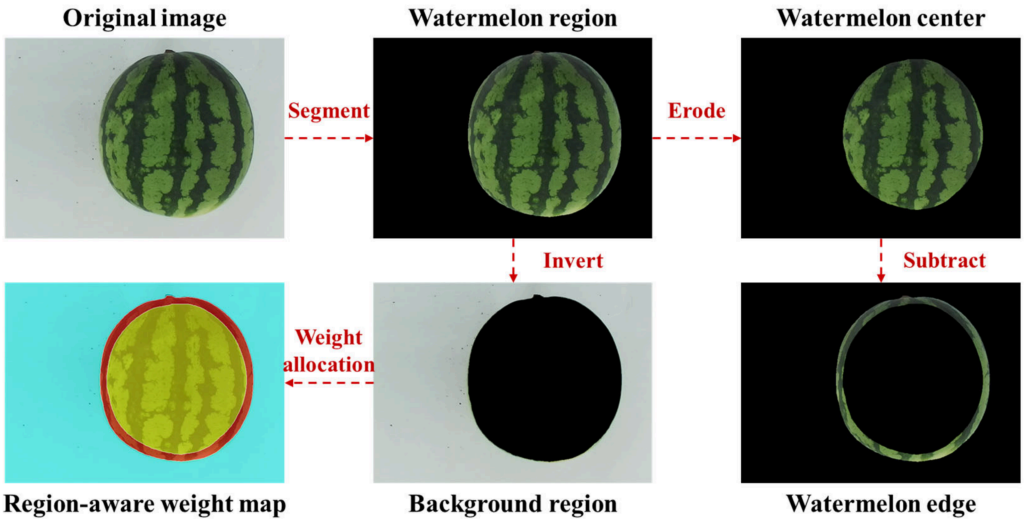


Figure 7. Generation process of the region-aware weight map.

The main branch aims to predict the disparity correction offset. Its supervision signal should be the actual residual between the disparity ground truth and the initial disparity. Thus, the loss function for this branch calculates the weighted difference between the predicted residual Δd and the ground truth residual $(d_{gt} - d_{init})$:

$$L_{resid} = \frac{1}{N} \sum_{(u,v)} W(u,v) \cdot \text{Smooth}_{L1}\left(\Delta d(u,v) - \left(d_{gt}(u,v) - d_{init}(u,v)\right)\right) \tag{8}$$

The final total loss L_{total} is defined as the weighted sum of the two aforementioned losses:

$$L_{total} = \gamma_1 \cdot L_{init} + \gamma_2 \cdot L_{resid} \tag{9}$$

2.5. Training Pipeline and Parameter Settings

Given the limited sample size of the real-world watermelon dataset, transfer learning is adopted in this study to prevent overfitting. First, the large-scale synthetic dataset SceneFlow [40] is utilized for pre-training. This dataset contains 39,824 stereo image pairs with a resolution of 960 × 540 and provides dense disparity ground truth, which helps the network learn general geometric priors for stereo matching. Pre-training is divided into two stages: In the first stage, the main branch is frozen, and only the attention branch is trained for 24 epochs to ensure the stability of the initial disparity guidance. In the second stage, all parameters are unfrozen, and the entire network is trained for another 24 epochs. In the fine-tuning stage, the pre-trained weights from SceneFlow are loaded, and the watermelon dataset constructed in this study is used for fine-tuning. The input image size is adjusted to 736 × 528 pixels via center cropping. During this stage, all parameters of the network are updated, and the fine-tuning epoch is set to 300.

All experiments are implemented based on the PyTorch deep learning framework. The Adam optimizer is adopted, with momentum parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to 1.0×10^{-3}, and a combined adjustment strategy of linear

warm-up and cosine annealing is employed to suppress gradient oscillations in the early training stage and guide the model to converge smoothly to the global optimum. For the joint loss function proposed in Section 2.4.4, the hyperparameters are set as $\gamma_1 = 0.5$ (weight of the attention branch) and $\gamma_2 = 1.0$ (weight of the main branch). Detailed parameter configurations are listed in Table 1.

Table 1. Training parameter settings for the MI-ACVNet watermelon stereo matching network.

Parameters	Value
Image size (pixels)	736×528
Batch size	12
Initial learning Rate	1.0×10^{-3}
Epochs	300
Max disparity	224

2.6. Evaluation Metrics

To quantitatively evaluate the accuracy and robustness of the MI-ACVNet model in the watermelon stereo matching task, three core metrics widely adopted in the stereo matching community are employed: End-Point Error (EPE), Root Mean Square Error (RMSE), and outlier percentage (Bad-x). All metrics are computed exclusively within valid pixel regions where disparity ground truth is available.

(1) End-Point Error (EPE)

EPE is the most fundamental metric for assessing stereo matching precision, defined as the average Euclidean distance between predicted and ground-truth disparities across all valid pixels. It reflects the overall pixel-level matching accuracy of the model, with smaller values indicating higher quality of the disparity map. The calculation formula is given as:

$$EPE = \frac{1}{N}\sum_{i\in\Omega}\left|d_{pred}^{(i)} - d_{gt}^{(i)}\right| \tag{10}$$

where N is the total number of valid pixels, Ω is the set of valid pixels, and $d_{pred}^{(i)}$ and $d_{gt}^{(i)}$ denote the predicted and ground-truth disparity values of the i-th pixel, respectively.

(2) Root Mean Square Error (RMSE)

RMSE quantifies the dispersion of prediction errors. Compared to EPE, RMSE is more sensitive to large-error samples (e.g., outlier points or mismatches at edges). This metric is introduced to evaluate the model's effectiveness in suppressing edge outliers. The calculation formula is given as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i\in\Omega}\left(d_{pred}^{(i)} - d_{gt}^{(i)}\right)^2} \tag{11}$$

(3) Outlier Percentage (Bad-x)

Bad-x represents the proportion of valid pixels whose absolute error exceeds a given threshold x. It reflects the model's reliability under different accuracy requirements, with lower values indicating fewer outliers. Three threshold levels are set in this study: Bad-2.0 (error > 2 pixels) measures the failure rate of coarse matching; Bad-1.0 (error > 1 pixel) serves as the standard evaluation metric for stereo matching tasks; and Bad-0.5 (error > 0.5 pixels) assesses the model's capability for fine sub-pixel matching. The calculation is as follows:

$$\text{Bad}-x = \frac{1}{N}\sum_{i\in\Omega}\mathbb{I}\left(\left|d_{pred}^{(i)} - d_{gt}^{(i)}\right| > x\right) \times 100\% \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function (1 if the condition is satisfied, 0 otherwise).

3. Results

3.1. Hyperparameter Optimization of the Loss Function

In Section 2.4.4, the SGRA-Loss was proposed to address the large matching error at watermelon edges by applying differential supervision weights to edge, central, and background regions (i.e., λ_{edge}, λ_{center}, and λ_{bg}). To determine the optimal configuration of these hyperparameters, a grid search experiment was conducted in this section based on the complete model integrated with the MSFE feature extraction module and the C2F-CRC cascaded refinement module. Considering that the center region occupies the majority of the watermelon image and exhibits relatively uniform textures, the center weight λ_{center} was fixed at 1.0 as a baseline. Building upon this, various weight combinations for the edge and background regions were evaluated. Specifically, λ_{edge} was tested at four intervals (1.0, 1.2, 1.4, and 1.6), and λ_{bg} was evaluated at four intervals (0.4, 0.6, 0.8, and 1.0). Although the background region lacks valid depth information of the watermelon, its weight was kept above zero to preserve a certain degree of supervision, thereby enhancing the model's generalization capability.

Table 2 presents the quantitative evaluation results for the watermelon region under each parameter configuration. The results indicate that, compared to the original equal-weight setup, moderately increasing the edge weight λ_{edge} consistently reduces key metrics such as EPE and Bad-0.5. This confirms that strengthened edge supervision facilitates the learning of sharper contour features, thereby mitigating edge matching errors. Notably, the configuration of $\lambda_{edge} = 1.4$, $\lambda_{center} = 1.0$, and $\lambda_{bg} = 0.8$ yields the best overall performance, achieving the lowest EPE, RMSE, Bad-0.5, and Bad-1.0, along with the second-lowest Bad-2.0. Based on these findings, the default hyperparameter configuration for the MI-ACVNet loss function is finalized as $\lambda_{edge} = 1.4$, $\lambda_{center} = 1.0$, and $\lambda_{bg} = 0.8$.

Table 2. Experimental results of loss function hyperparameter optimization.

Weight Assignment			Watermelon Region				
Watermelon Edge	**Watermelon Center**	**Background**	**EPE**	**RMSE**	**Bad-0.5 (%)**	**Bad-1.0 (%)**	**Bad-2.0 (%)**
1.0	1.0	1.0	0.099	0.539	1.322	0.295	0.096
1.2	1.0	0.4	0.093	0.536	1.173	0.291	0.099
		0.6	0.103	0.547	1.368	0.308	0.096
		0.8	0.095	0.542	1.274	0.292	**0.094**
1.4	1.0	0.4	0.096	0.540	1.265	0.309	0.101
		0.6	0.094	0.541	1.222	0.291	0.097
		0.8	**0.091**	**0.536**	**1.159**	**0.278**	0.096
1.6	1.0	0.4	0.098	0.542	1.326	0.313	0.099
		0.6	0.097	0.543	1.261	0.298	0.097
		0.8	0.097	0.536	1.286	0.297	0.095

Note: Bold values indicate the optimal results for each evaluation metric.

3.2. Ablation Study on MI-ACVNet Architecture

To verify the contributions of the proposed modules to the overall model performance, an ablation study was conducted on the watermelon stereo matching dataset by incrementally adding these modules to the Fast-ACVNet+ baseline. The effectiveness of the following three modules was primarily evaluated:

(1) MSFE: Epipolar-Enhanced Multi-Scale Stereo Feature Extraction module;
(2) C2F-CRC: Coarse-to-Fine Cascaded Residual Correction module;

(3) SGRA-Loss: Semantics-Guided Region-Aware Loss.

The experimental results are presented in Table 3. "All" denotes the error computed over all valid pixels in the image, while "Watermelon" denotes the error computed only within the watermelon region. Comparing Exp.1 (baseline) with Exp.2, introducing the MSFE module reduces the EPE over the whole image by 6.7% and significantly decreases Bad-0.5 by 23.1%. These results indicate that, through network re-parameterization and epipolar-enhanced attention, the MSFE module improves the model's ability to distinguish weak textures and high-frequency stripe patterns on the watermelon surface. Comparing Exp.1 with Exp.3, the C2F-CRC module reduces the overall EPE by 9.7% and the RMSE by 3.3%. Meanwhile, Bad-0.5, Bad-1.0, and Bad-2.0 decrease by 24.9%, 30.5%, and 23.0%, respectively. This improvement is achieved by replacing the discrete Top-*k* ranking in the original network with local window search and sub-pixel interpolation, enabling more accurate continuous disparity prediction. When both the MSFE and C2F-CRC modules are integrated (Exp.4), the model achieves further performance gains. Based on Exp.4, Exp.5 further introduces SGRA-Loss for adaptive supervision. Compared with Exp.4, the EPE in the watermelon region decreases by another 8.1%, while Bad-0.5 is reduced by 12.3%. These results demonstrate that SGRA-Loss strengthens supervision on difficult matching regions, such as object boundaries, during training and effectively suppresses matching errors.

Table 3. Ablation study on the matching accuracy of key modules in MI-ACVNet.

Exp.	MSFE	C2F-CRC	SGRA-Loss	Region	EPE	RMSE	Bad-0.5 (%)	Bad-1.0 (%)	Bad-2.0 (%)
Exp.1				All	0.134	0.649	2.489	0.380	0.126
				Watermelon	0.113	**0.526**	1.769	0.388	0.111
Exp.2	√			All	0.125	0.630	1.914	0.258	0.097
				Watermelon	0.106	0.549	1.470	0.329	0.105
Exp.3		√		All	0.121	0.628	1.870	0.261	0.097
				Watermelon	0.104	0.561	1.426	0.323	0.107
Exp.4	√	√		All	0.116	0.622	1.670	0.243	**0.091**
				Watermelon	0.099	0.539	1.322	0.295	0.096
Exp.5	√	√	√	All	**0.112**	**0.618**	**1.494**	**0.236**	0.093
				Watermelon	**0.091**	0.536	**1.159**	**0.278**	**0.096**

Note: "√" indicates that the corresponding module is utilized in the experiment; Bold values indicate the optimal results for each evaluation metric.

The effects of each module on computational complexity and efficiency are summarized in Table 4. Compared with Exp.1, the MSFE module in Exp.2 increases both the number of parameters (Params) and floating-point operations (FLOPs). This is mainly because the MobileOne backbone has more feature channels and branch structures than the original MobileNetV2. However, thanks to the re-parameterization mechanism of MobileOne during inference, the increase in model complexity has only a limited impact on runtime, and the inference time remains 47 ms. Comparing Exp.1 with Exp.3, the C2F-CRC module not only reduces the number of parameters but also shortens the inference time by 6.7%, reducing it to 28 ms. Comparing Exp.4 with Exp.5, SGRA-Loss is applied only during training. It introduces no additional parameters, FLOPs, or GPU memory consumption during inference, making it a zero-inference-cost optimization strategy.

Table 4. Ablation results of key MI-ACVNet modules on computational complexity and efficiency.

Exp.	MSFE	C2F-CRC	SGRA-Loss	Parameters (M)	FLOPs (G)	GPU Memory (MB)	Time (ms)
Exp.1				**3.20**	**64.24**	**419.10**	**30**
Exp.2	√			33.18	240.58	825.93	47
Exp.3		√		3.10	62.78	409.04	28
Exp.4	√	√		32.55	236.18	783.60	46
Exp.5	√	√	√	32.55	236.18	783.60	46

Note: "√" indicates that the corresponding module is utilized in the experiment; Bold values indicate the optimal results for each evaluation metric.

Overall, compared with the baseline model (Exp.1), the complete MI-ACVNet (Exp.5) achieves a cumulative reduction of 19.5% in EPE and 34.5% in Bad-0.5 in the watermelon region. This demonstrates that the proposed improvement strategies exhibit significant advantages in the 3D reconstruction task of watermelon surfaces, enabling a substantial reduction in stereo matching errors while ensuring real-time performance.

3.3. Comparison of Different Stereo Matching Algorithms

To further validate the overall advantages of MI-ACVNet for 3D reconstruction of watermelon surfaces, six representative lightweight stereo matching networks were selected for comparison with the proposed method. All experiments were conducted under the same hardware and software settings using the same dataset. Error evaluation was performed only within the watermelon region. The results are presented in Table 5. As shown in Table 5, MI-ACVNet achieves the best performance on almost all key accuracy metrics. The only exception is RMSE, where it ranks second, slightly behind Fast-ACVNet+. Compared with AANet [41], HSMNet [42], and LightStereo-L [34], MI-ACVNet reduces the EPE by 39.7%, 28.7%, and 53.3%, respectively, demonstrating clear performance advantages. Benefiting from the powerful feature extraction capability of the Transformer, NMRF-swint [43] achieves an EPE of 0.098 pixels, the best among the six compared methods. Even so, MI-ACVNet still outperforms it by 7.1% in EPE and 19.4% in Bad-0.5. These results further indicate that the proposed epipolar-enhanced attention, sub-pixel interpolation, and region-aware loss are more effective than a general Transformer architecture in capturing the subtle texture variations on watermelon surfaces.

Table 5. Performance comparison results of different stereo matching models.

Model	Watermelon Region								
	EPE	RMSE	Bad-0.5 (%)	Bad-1.0 (%)	Bad-2.0 (%)	Parameters (M)	FLOPs (G)	GPU Memory (MB)	Time (ms)
StereoNet	2.398	2.900	86.742	73.594	49.16	**0.42**	**45.75**	**157.28**	**17**
AANet	0.151	0.793	2.994	0.664	0.216	4.00	122.29	895.77	132
HSMNet	0.136	0.591	2.690	0.580	0.193	3.17	55.64	240.83	23
LightStereo-L	0.195	0.673	4.305	1.049	0.329	24.29	77.81	550.40	73
NMRF-swint	0.098	0.720	1.438	0.375	0.121	33.12	248.23	925.77	88
Fast-ACVNet+	0.113	**0.526**	1.769	0.388	0.111	3.20	64.24	419.10	30
MI-ACVNet (Ours)	**0.091**	0.536	**1.159**	**0.278**	**0.096**	32.55	236.18	783.60	46

Note: Bold values indicate the optimal results for each evaluation metric.

Table 5 also reports the model complexity and efficiency of all methods, including the number of parameters, FLOPs, GPU memory consumption, and inference time. MI-ACVNet has 32.55 M parameters and 236.18 G FLOPs, which are comparable to those of NMRF-swint (33.12 M parameters and 248.23 G FLOPs). Both models are larger than

StereoNet, HSMNet, and Fast-ACVNet+, which have fewer parameters. As shown in the ablation study (Table 4), the increase in model size mainly comes from replacing MobileNetV2 with the MobileOne backbone in the MSFE module. Notably, although MI-ACVNet has a model size and FLOPs similar to those of NMRF-swint, its inference time is only 46 ms, which is 47.7% shorter. This significant speed advantage benefits from the re-parameterization mechanism of MobileOne. During inference, the multi-branch structure used in training is equivalently merged into a single convolution branch, greatly reducing forward latency while maintaining representation capability. Compared with the baseline Fast-ACVNet+, MI-ACVNet increases the inference time by only 16 ms, while reducing the EPE by 19.5% and Bad-0.5 by 34.5%. This demonstrates a substantial improvement in accuracy at a modest cost in inference speed. In terms of GPU memory consumption, MI-ACVNet requires 783.60 MB. Although this is higher than that of some lightweight models, it still shows good potential for industrial deployment.

3.4. Three-Dimensional Point Cloud Reconstruction and Error Analysis

To intuitively evaluate the 3D reconstruction performance of different models, this section back-projects the disparity maps output by each model into 3D space based on camera intrinsic parameters and baseline distance to generate single-view watermelon point clouds. Meanwhile, the absolute depth errors of the predicted point clouds (relative to the ground truth along the Z-axis) are calculated and mapped as pseudo-color heatmaps on the point cloud surfaces. The visualization results are shown in Figure 8. In the figure, blue represents low errors, while red indicates high errors ($\geq$1.0 mm). Statistical analysis reveals significant differences in the average depth errors among models. Due to an excessively high downsampling rate, StereoNet yields an average depth error of up to 8.77 mm. The average depth errors of LightStereo-L, AANet, and HSMNet are 0.65 mm, 0.53 mm, and 0.40 mm, respectively. The well-performing NMRF-swint and the baseline model Fast-ACVNet+ achieve 0.27 mm and 0.33 mm, respectively. In contrast, the proposed MI-ACVNet attains the optimal reconstruction accuracy, with an average depth error of only 0.26 mm. Converted according to the binocular ranging principle, this depth error corresponds to a geometric dimension measurement error of approximately 0.09 mm for watermelons.

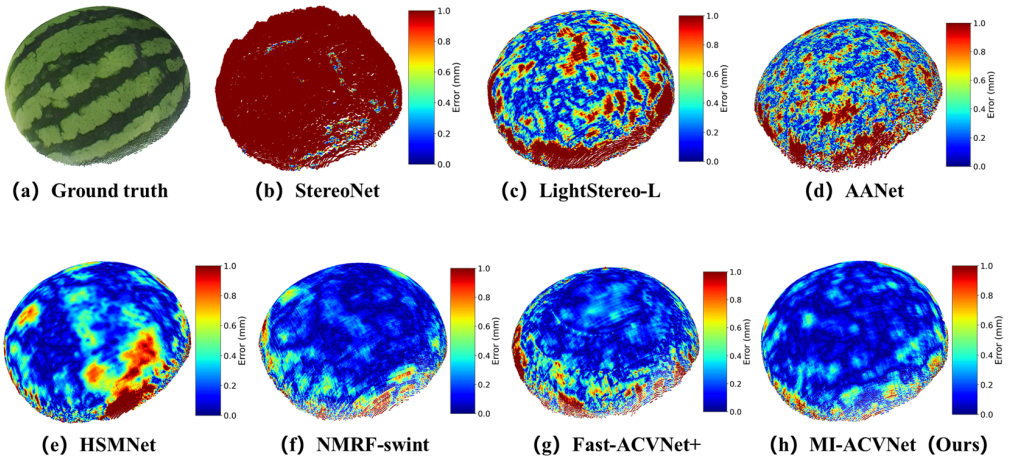


Figure 8. Error distribution of reconstructed point clouds from different models.

As observed from the figure, the reconstructed point clouds of nearly all models exhibit a trend of lower errors in the central region and higher errors at the edges. StereoNet, LightStereo-L, AANet, and HSMNet perform poorly, with large red or yellow regions distributed across their point cloud surfaces. In comparison, NMRF-swint, Fast-ACVNet+,

and MI-ACVNet deliver better reconstruction results, with their reconstructed surfaces dominated by blue, effectively restoring the geometric morphology of watermelons. Further comparison reveals that MI-ACVNet significantly outperforms other models in edge reconstruction. These results fully demonstrate that the multiple improvement strategies proposed in this study can specifically address the issue of low matching accuracy in weak-texture and edge regions on watermelon surfaces, substantially enhancing both global and local reconstruction quality while maintaining inference speed.

4. Discussion

The ultimate goal of the proposed single-view high-precision 3D reconstruction method is to provide technical support for automated watermelon grading. The major international grading standards generally classify watermelons based on three categories of geometric indicators: fruit size (diameter and weight), shape regularity (degree of deformity), and surface defects (scars, cracks, sunburn, etc.). These standards include the United States Standards for Grades of Watermelons established by the USDA Agricultural Marketing Service (USDA AMS), the UNECE Standard FFV-37 Concerning the Marketing and Commercial Quality Control of Watermelons, and the Chinese national standard for seedless watermelon grading. Watermelons are typically classified into Extra Class, Class I, and Class II. Extra Class requires a uniform shape and no surface defects. Class I allows slight shape deviations and very minor surface defects, with a total defect area of less than 5 cm^2. Class II meets the minimum market requirements but permits larger defects, with a total defect area of less than 10 cm^2. These standards do not specify explicit accuracy requirements for grading. Existing studies generally report dimensional measurement errors within 1–3 mm for fruits [9,14]. It should be noted that the current single-view point cloud only covers the top visible surface of the watermelon. The measurement of grading indicators such as diameter, volume, and full-surface defect area requires a complete 3D watermelon model. Building such a model, in turn, relies on multi-view point cloud registration and filling of the occluded bottom region. The orthogonal multi-view camera array shown in Figure 1 is designed for full-surface reconstruction. However, this study focuses solely on improving the accuracy of single-view stereo matching. This is because single-view point clouds serve as the foundational building block for subsequent multi-view fusion, their accuracy directly determines the upper bound of the final full-surface reconstruction quality. The average depth error of 0.26 mm achieved in this work (corresponding to a geometric measurement error of approximately 0.09 mm for a typical watermelon with a diameter of about 20 cm) provides a solid accuracy foundation for subsequent multi-view full-surface reconstruction and precise measurement of grading indicators.

In terms of generalization, the key improvements of MI-ACVNet are not designed for a specific watermelon variety. Instead, they target the common characteristics of near-spherical fruits and agricultural products. The three proposed strategies, namely Epipolar-Enhanced Coordinate Attention (E2CA), Coarse-to-Fine Cascaded Residual Correction (C2F-CRC), and Semantics-Guided Region-Aware Loss (SGRA-Loss), are all designed based on these general geometric and textural characteristics. Therefore, the proposed method is applicable to other near-spherical fruits and agricultural products with similar properties, such as muskmelons, cantaloupes, citrus fruits, tomatoes, apples, and even some tuber crops. For a new target category, only stereo images and ground-truth data need to be collected for fine-tuning, enabling rapid deployment.

It should be noted that all experiments in this study were conducted under a single controlled lighting condition using a cylindrical light box. This setup was not intended to simplify the experiments. Instead, it was designed to simulate the actual imaging environment inside industrial fruit sorting systems, where engineered lighting is widely used to

ensure uniform and stable illumination. However, the current experiments do not cover more challenging lighting conditions, such as strong light, weak light, and non-uniform illumination. The robustness of MI-ACVNet under these conditions still requires further investigation. Evaluating the proposed method under multiple lighting conditions will be an important direction for future work. In addition, the stereo matching dataset constructed in this study contains 808 stereo image pairs collected from 101 Kirin watermelons. The samples exhibit considerable variation in size, shape, and stripe distribution. Nevertheless, the current dataset still covers a limited range of varieties, production regions, and lighting conditions. As a result, the model performance may degrade when applied to watermelon varieties with weak or no surface texture or under complex lighting conditions. Expanding the dataset to include more varieties, production regions, and illumination conditions will further improve the generalization ability and practical applicability of the proposed method.

5. Conclusions

To achieve high-precision single-view point cloud acquisition of Kirin watermelons, a deep learning stereo matching network, MI-ACVNet, is proposed. Built upon the Fast-ACVNet+ baseline, the network integrates an Epipolar-Enhanced MSFE module, a C2F-CRC strategy, and a SGRA-Loss. Ablation studies demonstrate that each proposed module yields significant performance gains. Compared to the baseline model, the full MI-ACVNet reduces the EPE by 19.47% and the Bad-0.5 error rate by 34.48% in the watermelon region, achieving an inference time of only 46 ms. Furthermore, when compared against mainstream algorithms such as StereoNet, AANet, and NMRF-swint, MI-ACVNet achieves the optimal reconstruction performance while ensuring real-time efficiency, yielding an average depth error of only 0.26 mm. This method effectively overcomes the matching challenges posed by the highly similar textures and large edge curvatures of watermelons, providing robust technical support for the external quality inspection and geometric feature evaluation of watermelons.

Author Contributions: Z.L.: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing—original draft. X.X.: Formal analysis, Data curation, Writing—review and editing. Y.G.: Formal analysis, Data curation, Writing—review and editing. W.L.: Data curation. X.R.: Conceptualization, Funding acquisition, Supervision, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (Grant number:2022YFD2002200).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: During the preparation of this manuscript, the authors used Doubao (developed by ByteDance, Beijing, China, online version, continuously updated) for the purposes of translation and language polishing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kenea, F.T.; He, N.; Lu, X.; Luo, X.; Zhu, H.; Liu, W. Watermelon Fruit Metabolome Gene Discovery and Its Application in Breeding: A Review. *Front. Plant Sci.* **2025**, *16*, 1687406. [CrossRef] [PubMed]
2. Gidado, M.J.; Gunny, A.A.N.; Gopinath, S.C.; Devi, M.; Ibrahim, A.M. Artificial Intelligence and Machine Learning in Postharvest Fruit Quality Assessment: Current Challenges, Recent Advances, and Future Prospects. *J. Stored Prod. Res.* **2026**, *116*, 102959. [CrossRef]
3. Yu, G.; Ma, B.; Li, Y.; Dong, F. Quality Detection of Watermelons and Muskmelons Using Innovative Nondestructive Techniques: A Comprehensive Review of Novel Trends and Applications. *Food Control* **2024**, *165*, 110688. [CrossRef]
4. Liu, J.; Sun, J.; Wang, Y.; Liu, X.; Zhang, Y.; Fu, H. Non-Destructive Detection of Fruit Quality: Technologies, Applications and Prospects. *Foods* **2025**, *14*, 2137. [CrossRef] [PubMed]
5. Wang, R.-F.; Xu, R.; Chee, P.W.; Wang, H.; Li, C. A Review of Visual Navigation for Agricultural Robots in Open Fields and Controlled Environments. *Comput. Electron. Agric.* **2026**, *248*, 111754. [CrossRef]
6. Wang, R.-F.; Qu, H.-R.; Su, W.-H. From Sensors to Insights: Technological Trends in Image-Based High-Throughput Plant Phenotyping. *Smart Agric. Technol.* **2025**, *12*, 101257. [CrossRef]
7. Zhou, Z.; Zahid, U.; Majeed, Y.; Nisha; Mustafa, S.; Sajjad, M.M.; Butt, H.D.; Fu, L. Advancement in Artificial Intelligence for On-Farm Fruit Sorting and Transportation. *Front. Plant Sci.* **2023**, *14*, 1082860. [CrossRef] [PubMed]
8. Koc, A.B. Determination of Watermelon Volume Using Ellipsoid Approximation and Image Processing. *Postharvest Biol. Technol.* **2007**, *45*, 366–371. [CrossRef]
9. Hong, S.-J.; Kim, J.; Lee, A. Real-Time Morphological Measurement of Oriental Melon Fruit through Multi-Depth Camera Three-Dimensional Reconstruction. *Food Bioprocess Technol.* **2024**, *17*, 5038–5052. [CrossRef]
10. Yamamoto, S.; Karkee, M.; Kobayashi, Y.; Nakayama, N.; Tsubota, S.; Thanh, L.N.T.; Konya, T. 3D Reconstruction of Apple Fruits Using Consumer-Grade RGB-Depth Sensor. *Eng. Agric. Environ. Food* **2018**, *11*, 159–168.
11. Wang, Y.; Chen, Y. Fruit Morphological Measurement Based on Three-Dimensional Reconstruction. *Agronomy* **2020**, *10*, 455. [CrossRef]
12. Neupane, C.; Koirala, A.; Wang, Z.; Walsh, K.B. Evaluation of Depth Cameras for Use in Fruit Localization and Sizing: Finding a Successor to Kinect V2. *Agronomy* **2021**, *11*, 1780. [CrossRef]
13. Masoudi, M.; Golzarian, M.R.; Lawson, S.S.; Rahimi, M.; Islam, S.M.S.; Khodabakhshian, R. Improving 3D Reconstruction for Accurate Measurement of Appearance Characteristics in Shiny Fruits Using Post-Harvest Particle Film: A Case Study on Tomatoes. *Comput. Electron. Agric.* **2024**, *224*, 109141. [CrossRef]
14. Ma, H.; Zhu, X.; Ji, J.; Wang, H.; Jin, X.; Zhao, K. College of Agricultural Equipment Engineering, Henan University of Science and Technology, Luoyang 471003, Henan, China Rapid Estimation of Apple Phenotypic Parameters Based on 3D Reconstruction. *Int. J. Agric. Biol. Eng.* **2021**, *14*, 180–188. [CrossRef]
15. Wu, F.; Chen, Y. Fruitninja: 3D Object Interior Texture Generation with Gaussian Splatting. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 11051–11060.
16. Yi, W.; Xia, S.; Kuzmin, S.; Gerasimov, I.; Cheng, X. RTFVE-YOLOv9: Real-Time Fruit Volume Estimation Model Integrating YOLOv9 and Binocular Stereo Vision. *Comput. Electron. Agric.* **2025**, *236*, 110401.
17. Mirbod, O.; Choi, D.; Heinemann, P.H.; Marini, R.P.; He, L. On-Tree Apple Fruit Size Estimation Using Stereo Vision with Deep Learning-Based Occlusion Handling. *Biosyst. Eng.* **2023**, *226*, 27–42.
18. Gao, Y.; Wang, Q.; Rao, X.; Xie, L.; Ying, Y. OrangeStereo: A Navel Orange Stereo Matching Network for 3D Surface Reconstruction. *Comput. Electron. Agric.* **2024**, *217*, 108626. [CrossRef]
19. Liu, W.; Wang, C.; Yan, D.; Chen, W.; Luo, L. Estimation of Characteristic Parameters of Grape Clusters Based on Point Cloud Data. *Front. Plant Sci.* **2022**, *13*, 885167. [CrossRef] [PubMed]
20. Wasenmüller, O.; Stricker, D. Comparison of Kinect V1 and V2 Depth Images in Terms of Accuracy and Precision. In *Computer Vision—ACCV 2016 Workshops*; Chen, C.-S., Lu, J., Ma, K.-K., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2017; Volume 10117, pp. 34–45. ISBN 978-3-319-54426-7.
21. Dashpute, A.; Anand, C.; Sarkar, M. Depth Resolution Enhancement in Time-of-Flight Cameras Using Polarization State of the Reflected Light. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 160–168.
22. Meinen, B.U.; Robinson, D.T. Mapping Erosion and Deposition in an Agricultural Landscape: Optimization of UAV Image Acquisition Schemes for SfM-MVS. *Remote Sens. Environ.* **2020**, *239*, 111666. [CrossRef]
23. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *Commun. ACM* **2022**, *65*, 99–106. [CrossRef]
24. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.* **2023**, *42*, 1–14. [CrossRef]
25. Yao, M.; Huo, Y.; Ran, Y.; Tian, Q.; Wang, R.; Wang, H. Neural Radiance Field-Based Visual Rendering: A Comprehensive Review. *IEEE Trans. Vis. Comput. Graph.* **2026**, *32*, 7361–7379. [CrossRef]

26. Li, Z.; Xu, X.; Xu, T.; Gao, Y.; Rao, X. Process, Challenges and Solutions of Fruit 3D Reconstruction: A Review. *Comput. Electron. Agric.* **2025**, *239*, 111068. [CrossRef]
27. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341.
28. Chang, J.-R.; Chen, Y.-S. Pyramid Stereo Matching Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 3–7 June 2018; pp. 5410–5418.
29. Xu, G.; Cheng, J.; Guo, P.; Yang, X. Attention Concatenation Volume for Accurate and Efficient Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12981–12990.
30. Chen, Z.; Long, W.; Yao, H.; Zhang, Y.; Wang, B.; Qin, Y.; Wu, J. Mocha-Stereo: Motif Channel Attention Network for Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 27768–27777.
31. Ma, R.; Pan, F.; Liang, P.; Diao, L.; Li, J.; Zhang, B.; Qi, L. Real2Real: Invariant Style Alignment for Unsupervised One-Shot Real Noisy Image Denoising. *Inf. Fusion* **2026**, *135*, 104410. [CrossRef]
32. Khamis, S.; Fanello, S.; Rhemann, C.; Kowdle, A.; Valentin, J.; Izadi, S. Stereonet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 573–590.
33. Xu, G.; Wang, Y.; Cheng, J.; Tang, J.; Yang, X. Accurate and Efficient Stereo Matching via Attention Concatenation Volume. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 2461–2474. [CrossRef]
34. Guo, X.; Zhang, C.; Zhang, Y.; Zheng, W.; Nie, D.; Poggi, M.; Chen, L. Lightstereo: Channel Boost Is All You Need for Efficient 2d Cost Aggregation. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19–23 May 2025; pp. 8738–8744.
35. Vasu, P.K.A.; Gabriel, J.; Zhu, J.; Tuzel, O.; Ranjan, A. Mobileone: An Improved One Millisecond Mobile Backbone. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 7907–7917.
36. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
37. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
38. Lipson, L.; Teed, Z.; Deng, J. Raft-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; pp. 218–227.
39. Khanam, R.; Hussain, M. Yolov11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725.
40. Mayer, N.; Ilg, E.; Hausser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4040–4048.
41. Xu, H.; Zhang, J. Aanet: Adaptive Aggregation Network for Efficient Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 1959–1968.
42. Yang, G.; Manela, J.; Happold, M.; Ramanan, D. Hierarchical Deep Stereo Matching on High-Resolution Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 5515–5524.
43. Guan, T.; Wang, C.; Liu, Y.-H. Neural Markov Random Field for Stereo Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 5459–5469.