

Depth Map Reconstruction in Underwater Robotic Vision using Gradient Spanning Trees and Boundary Segmentation

Nurliyana Abd Mutalib¹, Rostam Affendi Hamzah^{2*}, Ken Prameswari Caesarella Aryaputri³, Ahmad Fauzan Kadmin⁴, Masrullizam Mat Ibrahim⁴, and Adi Irwan Herman⁵

^{*1,2,3}Engineering Technology Department, Universiti Teknikal Malaysia Melaka, 76100 Melaka, Malaysia

⁴Engineering Department, Universiti Teknikal Malaysia Melaka, 76100 Melaka, Malaysia

⁵Texas Instruments Sdn Bhd, 74000 Melaka, Malaysia

*Correspondence: rostamaffendi@utem.edu.my

ABSTRACT- Underwater robotic vision robot encounters multiple challenges. Light is absorbed and scattered resulting in depth perception difficulties. The complexity of underwater scenes is worsened due to low texture, plain color, and irregular object boundaries. Hence, this article proposes a new algorithm to reconstruct disparity maps or depth maps from a stereo vision system used for underwater robotic vehicles. Fundamentally, there are four stages to develop the algorithm which are matching cost computation, cost aggregation, depth selection and depth refinement. To initiate the process a preliminary depth map is obtained using a bit-wise census-based matching cost computation. A high efficiency level is shown by this method even with differing lighting or colors on images. Cost aggregation employs a weighting approach from gradient Spanning Tree (ST) which applies gradient variations for enhancing the initial depth map. This method integrates gradient information into the spanning tree construction allowing the algorithm to preserve boundary discontinuities more effectively, reducing the risk of smoothing across object edges, unlike the common ST approaches that rely solely on edge weights. Next stage is depth selection which adopts a "winner-takes-all" (WTA) strategy. The depth value is determined by selecting minimum cost volume from normalization process. The final stage, boundary-based segmentation is proposed to refine final depth map. This method reconstructs ill-posed depth maps and increases the accuracy. Additionally, textures are added to object edges during this step resulting in precision of the final depth maps. Unlike prior ST-based and segmentation approaches, the proposed method uniquely integrates gradient-weighted spanning tree aggregation with boundary segmentation refinement, specifically proposed to overcome underwater vision challenges and complexities. Evaluations of the standard Middlebury system attains a non-occ error of 7.59%, all error of 10.90% and end-point-error at 0.5-1.0 demonstrating competitive performance compared with existing methods such as Esmea, PPEP-GF, Dense-CNN, ACMC and RDNet. The real images from the underwater Caddy benchmarking dataset further demonstrate that the proposed method produces clearer and more visually accurate results.

Keywords: Boundary Segmentation, Depth Map, Gradient, Spanning Tree, Underwater Vision.

ARTICLE INFORMATION

Author(s): Nurliyana Abd Mutalib, Rostam Affendi Hamzah, Ken Prameswari Caesarella Aryaputri, Ahmad Fauzan Kadmin, Masrullizam Mat Ibrahim, and Adi Irwan Herman;

Received: 23/05/26; **Accepted:** 03/09/26; **Published:** 30/09/26;

E- ISSN: 2347-470X;

Paper Id: IJEER 2305B05;

Citation: 10.37391/ijeer.140316

Webpage-link:

<https://ijeer.forexjournal.co.in/archive/volume-14/ijeer-140316.html>



Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

Stereo vision plays a key role in underwater vision system. It allows any applications to perceive depth and 3D shapes by using just two images. This ability proves helpful in various fields such as robotics, self-driving vehicles or autonomous vehicles, augmented reality and machine vision applications in factories [1],[2],[3]. These applications need accurate and prompt depth data to work well. Fundamentally, stereo

matching corresponds pixels from two images. It produces depth maps that contain information of depth and indicate the distance of objects. While stereo matching algorithms have advanced significantly now a days, it still faces challenges. It has a challenging process with regions that lack textures, plain color regions and discontinuity or quick shifts in depth or lighting. These challenges reduce the accuracy of depth map and depth perception. Basically, stereo matching algorithms have four stages which are matching cost computation, cost aggregation, depth selection and depth refinement [4].

Matching cost computation is the first stage which generates preliminary depth map. A common pixel-based method for matching cost computation is absolute differences [5][6]. This technique produces high errors in complex regions and is sensitive to radiometric differences. Additionally, block based matching cost techniques such as Sum of Absolute Differences (SAD) and Sum of Squared Differences (SSD) are sensitive to illumination differences and texture-less surfaces [7]. Another prominent method at this stage is Census transform. This technique converts local brightness patterns into easy binary

codes. It performs effectively when there are significant light changes between images. It struggles in places with low details or when some parts are obscured [8][9][10]. To address these, researchers have worked on creating improved versions of Census and applied adaptive windows for better results. Merging Census with other techniques has led to better outcomes in challenging regions such implemented by [11][12].

The second step, known as cost aggregation holds a significant importance in stereo matching framework. Edges in the depth map may be blurred by traditional local methods [13][14][15]. Furthermore, global methods such as belief propagation [17] and graph cuts [18] do not guarantee superior quality without requiring substantial computing power. Researchers have explored non-local aggregation techniques to achieve balance [19][20][21]. Aggregation methods based on MST are considered effective due to their efficient distribution of costs across interconnected areas [22][23]. Maintaining accuracy and speed is aided by this approach. Alternative techniques incorporate gradient-based fusion or models for cross-scale costs to enhance depth maps. Results are often improved by these methods in regions characterized by sparse texture or proximity to occlusions. Such methods do not guarantee improvements in all cases especially when texture is abundant [24][25][26]. *Table 1* provides a comparison of several segment tree and tree-based methods applied to cost aggregation. The results indicate that existing ST-based techniques exhibit certain strengths and limitations. Generally, existing methods face a trade-off between boundary preservation, robustness and model complexity.

Table 1. Comparison table with current available and proposed methods based on the Segment Tree

Algorithm	Approach	Limitations
Standard Spanning Tree (ST) Aggregation [10]	Aggregates disparity costs along tree edges	Relies solely on pixel similarity; ignores structural boundaries
Minimum Spanning Tree (MST) [16]	Builds spanning tree using edge weights (intensity/color similarity)	Tends to over smooth object boundaries; weak in occluded regions
Segmentation-based ST Variants [18]	Use region segmentation to guide aggregation	Segmentation errors propagate; limited adaptability to fine gradients
Dense CNN / Deep ST Hybrids [36]	Learn cost aggregation via deep features	Computational expensive; less interpretable

Depth selection typically follows a winner-takes-all (WTA) approach [8][19]. The lowest cost is utilized to determine depth values. The advantage of WTA is reliability of the depth map that maintains quick computation. Ensuring it is always reliable is a consideration to use this strategy. At the final stage, depth refinement should increase the accuracy of final depth map. In recent years, segmentation-based refinement has gained significant advancement [23][27]. Superpixel segmentation

[28], hierarchical models [29][30] and segmentation-guided matching [31] are utilized for maintaining object edges while ensuring smoothness within inner regions. It's not uncommon for these methods to enhance visual quality and coherence. Segment-based method proves effective for specific image categories such as satellite stereo images and expansive natural landscapes [32][33]. Accuracy is enhanced through the refinement of depth maps with a focus on regions leading to a decrease in errors in complex regions. Errors in tough parts are not eliminated by these methods.

Improvements in older techniques have been made by incorporating color clues, edge information and filters for clarity [3][8][15]. Better depth maps are achieved when these different techniques are utilized. Not all methods lead to improved results in depth maps. Complex techniques utilize algorithms that are aware of context [34][35]. Complicated areas are addressed through the fixing of depth maps by these algorithms. Such methods do not only work effectively in simple regions. Researchers generally agree that the optimal performance of stereo matching arises from a combination of various methods [23][26]. Strong features are represented through local descriptors which are employed alongside non-local combination to disseminate information effectively. It is true that the structures remain improper without graph-based refinement [6].

In recent years stereo matching has undergone significant transformation due to deep learning advancement. High accuracy has been achieved in popular evaluations such as KITTI and Middlebury [27] through end-to-end convolutional neural networks (CNNs) and cost volume networks. Most traditional methods do not perform as well. Accuracy is dominated by deep learning techniques [29][36]. Deep learning presents certain drawbacks. A substantial amount of labeled data is required for the training process. It fails to perform effectively with unfamiliar image types. High computing power and memory resources are also necessary for proper functioning of deep learning systems [3][6][14]. As a result, traditional or non-deep learning methods are always inferior for real-time applications and compact devices. These methods operate efficiently without the need for specialized hardware enabling quicker performance. Numerous researchers believe that effective classical techniques play a crucial role in achieving a balance between precision efficiency and user-friendliness in practical scenarios. While ST-based aggregation [12], [15], [18], and segmentation methods [1], [21], [32] have been explored in stereo vision, their application to underwater robotic vision remains limited. The proposed contribution lies in combining gradient-weighted spanning tree cost propagation with boundary segmentation, enabling sharper edge preservation and improved robustness in texture-poor underwater environments.

Over the last few years, underwater stereo vision research has adopted traditional pinhole camera methods to model the optical physics of subsea environments. This method fails due

to light refraction at the air–glass–water boundaries such implemented by [38]. Additionally, they used virtual optical focal shifts and non-linear ray transformations to neutralize scale drift caused by refraction before stereo matching occurs. Deep unsupervised stereo matching uses semantic attention and cross-view transformer modules such implemented by [39] to overcome low-contrast and textureless underwater scenes. By incorporating high-level visual categories into the cost-aggregation stage, this method maintains matching accuracy despite heavy backscatter and producing some noise on repetitive regions. Another approach by [40] known as *UWStereo* paired with active structured-light setups to supply dense ground-truth disparity maps. This method has substantially improved disparity estimation accuracy for non-invasive marine biometric measurement, subsea pipeline inspection, and autonomous underwater vehicle (AUV) navigation. However, this method still produces some errors in disparity map results especially on plain color regions and the use of structured light requires high energy which is not suitable for underwater vision applications.

Therefore, this article proposes a new stereo matching algorithm based on census transform at first stage. Changes in lighting and radiometric error are managed effectively by this method. Then in the second stage, Spanning Tree (ST) gradient weighted differences are proposed to rapidly and smoothly disseminated across plain color regions. Boundary segmentation is introduced at final stage to maintain depth edges clearly while creating a smooth final depth map. This novel approach addresses frequent issues encountered in complex regions including reconstruction of final depth map. Challenges such as plain color regions, lack of texture and abrupt depth transitions or depth discontinuity are resolved effectively. Hence, these points are the contributions:

A novel stereo vision-based depth map reconstruction algorithm is designed for vision-based applications.

A new cost aggregation method is proposed using gradient spanning tree to improve edge preservation.

A boundary-based segmentation is developed at final stage by reconstructing ill-posed regions leading to improved final depth map.

2. MATERIALS AND METHODS

Stereo vision estimates depth by computing a depth map, which translates the horizontal displacement of pixel correspondences between rectified image pairs. Fundamentally, Figure 1 displays a stereo matching algorithm. It involves four stages which are matching cost computation, cost aggregation, disparity selection, and disparity refinement. The reliability of each stage determines the accuracy of the final depth map. Therefore, the proposed work in this article at cost computation is utilizing census transformation. Local intensity structures are encoded into binary strings by Census, allowing for reliable matching even in the presence of radiometric errors and monotonic illumination variations. It maintains sensitivity to local texture

while guaranteeing computing efficiency through the use of Hamming distance. Next, aggregation based on weightage of edge gradient differences at ST replaces fixed windows with undirected graph segmentation, propagating costs along intensity-consistent edges. This ensures smooth distribution inside homogeneous regions while maintaining sharp object borders. At final stage, while maintaining inter-region discontinuities, boundary-based segmentation ensures intra-region depth smoothness. In order to rectify mismatches and sharpen structural edges, boundary segmentation aligns depth refinement with perceptually consistent regions. The effectiveness of this new algorithm is validated in results section through experimental evaluation using a standard Middlebury benchmark stereo evaluation system [27].

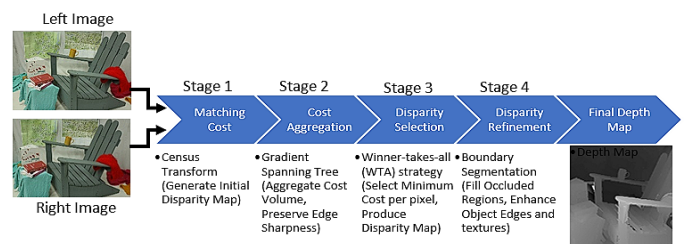


Figure 1. The proposed algorithm has four stages of framework

2.1. Matching Cost Computation

In this work, matching cost is computed using the Census Transform, a widely adopted non-parametric descriptor known for its robustness to radiometric distortions, illumination inconsistencies, and local noise [2],[10]. The Census Transform encodes local image structure by comparing the intensity of a pixel of interest $I(p)$ with each neighboring pixel $I(q)$ within a predefined window $N(p)$. Each function is represented by a binary equation (1) and equation (2) as follows:

$$\xi(I(p), I(q)) = \begin{cases} 1, & \text{if } I(q) < I(p) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The binary outcomes from all neighbors are concatenated using the bitwise operator $\oplus$, yielding the Census descriptor for pixel p in equation (2):

$$C(p, q) = \oplus_{q \in N(p)} \xi(I(p), I(q)) \quad (2)$$

where $I(p)$ and $I(q)$ denote the pixel of interest and neighbourhood pixel intensities respectively, with $N(p)$ defines its neighborhood within the selected q . In this article, a 9×9 window is adopted based on its validated effectiveness in prior research [2]. Additionally, this census window size was selected because it delivers adequate neighbourhood information to produce a discriminative descriptor. At this point, it maintains robustness against image noise and illumination variations. Compared with smaller windows, the 9×9 configuration improves matching reliability, whereas larger windows tend to oversmoothed depth discontinuities and increase computational complexity. Therefore, the 9×9 window offers an effective trade-off between accuracy, robustness, and processing efficiency. For stereo correspondence, descriptors from left and

right images are compared using Hamming distance, which counts the number of differing bits [31]. This approach effectively captures local structural variations, enabling reliable matching even in low-texture or unevenly lit regions [5],[14].

2.2. Cost Aggregation

The task at this stage is to refine the initial cost volume from matching cost process. This article proposes a new technique at this stage based on undirected graph segmentation from $(G = V, E)$ which uses a weightage of image gradient operation as given by equation (3):

$$G = (V, E, w(e)) \quad (3)$$

where V , E and w represent a set of vertices (*i.e.*, all the image pixels), edges (*i.e.*, all the edges between the nearest neighboring pixels) and weight function of image gradient respectively. This weight is represented by equation (4):

$$w(e) = \exp \left(-\frac{(I_p - I_q)}{\sigma} \right) \quad (4)$$

where σ is a constant parameter to adjust the similarity between two nodes, $(I_p - I_q)$ denotes the gradient difference pixel of interest p and neighbouring pixels q . The weight function $w(e)$ reflects spatial distance, color similarity, or edge affinity, ensuring that aggregation respects object boundaries and propagates information effectively. Then, the tree structure or parent node $T \subseteq E$ is reconstructed which contains of n and must satisfy the following equation (5):

$$T_{n+1} = T_n \cup \left\{ \min_{e \in T_n} w(e) \right\} \quad (5)$$

where e is the minimum edge across some cut and growing tree structure T_{n+1} which utilizes the Kruskal algorithm. At this stage, a new function represented by equation (6):

$$ST = \arg \min_{T \subseteq E} \sum_{e \in T} w(e) \quad (6)$$

where ST is tree structure to reconstruct the edge and segmented region. In this work, the tree structure is constructed with edge weights defined by pixel-wise gradient differences, enabling cost propagation along the tree while preserving object boundaries. This non-local design ensures that structural coherence is maintained and allows depth costs to be effectively propagated across semantically similar regions. Consequently, the aggregation process achieves enhanced robustness in low texture, plain color and depth discontinuity areas where traditional local support windows are insufficient to entertain these regions. The final aggregated cost at this stage is expressed as equation (7):

$$C_A(p, q) = C(p, q) + \arg \min_{T \subseteq E} \sum_{e \in T} \exp \left(\frac{I_p - I_q}{\sigma} \right) \quad (7)$$

This formulation supports long-range consistency, reduces noise in homogeneous areas, and improves depth accuracy in structurally complex regions or scenes.

2.3. Depth Selection

This stage employs the Winner-Takes-All (WTA) strategy, which remains the most widely used approach for assigning depth values due to its computational efficiency and straightforward implementation [32]. In WTA, each pixel is assigned the depth that minimizes the aggregated matching cost over the search range, as in equation (8):

$$d(p) = \arg \min_{d \in D} C_A(p, d) \quad (8)$$

where $d(p)$ denotes the selected depth for pixel p , D represents maximum value for depth and $C_A(p, d)$ is the aggregated cost obtained from the previous stage. This ensures that depth estimation is both fast and suitable for real-time [24].

2.4. Depth Refinement

In this article, refinement of the final depth map is achieved through a boundary-based segmentation, designed to improve depth accuracy near object boundaries and in low texture or discontinuity regions. The image is partitioned into coherent regions (boundary segment), with each region represented as a node in the depth map. Edges between regions were weighted by photometric similarity and spatial proximity. This forms a sparse and structured segmentation. Hence, a new boundary-based segmentation method is introduced at this stage to increase the final depth map accuracy. The edge (*i.e.*, v_p, v_q) is selected using the weight on each edge of the dissimilarity measure w_{ed} between the depth values as shown by Equation (9).

$$w_{ed}(v_p, v_q) = |d(p) - d(q)| \quad (9)$$

where v_p and v_q denoted as a vertice of interest and neighbour vertices respectively. The $d(p)$ and $d(q)$ are pixel of interest and neighbour depth pixels from WTA. Each segment has two conditions as follows. First, edges between depth values in the same component should have low weights, whereas those between pixels in different components should have greater weights. These two different components generate a link or boundary. Second, this boundary of interest B_p generates a segmented region thru the connected boundaries B_q and form a region as stated by equation (10) and equation (11).

$$\Delta(B_p, B_q) = \min_{v_p \in B_p, v_q \in B_q} w_{ed}(v_p, v_q) \quad (10)$$

$$S(B_p, B_q) = \begin{cases} B_p \cup B_q & \text{if } \Delta(B_p, B_q) < \Delta_{\max}(B_p, B_q) \\ (B_p \cup B_q)^c & \text{Disjoint} \end{cases} \quad (11)$$

where B_p and B_q merge as an union segment once the $\Delta(B_p, B_q)$ is lower than $\Delta_{\max}(B_p, B_q)$, otherwise the B_p and B_q may remain as a different segment (*i.e.*, disjoint). The $\Delta_{\max}(B_p, B_q)$ is given by Equation (12):

$$\begin{aligned} \Delta_{\max}(B_p, B_q) \\ = \max(\tau_{seg}(B_p), \tau_{seg}(B_q)) \end{aligned} \quad (12)$$

where τ_{seg} is a threshold to adapt according to the size of the segment as given by Equation (13):

$$\tau_{seg}(B) \quad (13)$$

where k is a constant parameter as a scale of observation and z_B represents the boundary size. The k controls behaviour of overall segmentation process which a larger value resulting in more aggressive merging while a smaller value producing stricter segmentation with preserving more boundaries. Therefore, this constant parameter determines the trade-off between disparity smoothing and edge preservation. Based on this proposed method, the resulting boundary will be broken down into simpler segments, which helps to reduce its complexity and simplify its description with refine boundaries. Hence, the pseudocode at this stage is given below to illustrate the flow of the depth refinement stage.

Input: Disparity map $d(p)$

Output: Refined disparity map $S(B_p, B_q)$

Compute edge weight:

$$w_{ed}(v_p, v_q) = |d(p) - d(q)|$$

Initialize each pixel as a segment

For each edge (v_p, v_q) sorted by $w_{ed} d$

$$\Delta(B_p, B_q) = \min_{v_p \in B_p, v_q \in B_q} w_{ed}(v_p, v_q)$$

$$\Delta_{max}(B_p, B_q) = \max(\tau_{seg}(B_p), \tau_{seg}(B_q))$$

If $\Delta(B_p, B_q) < \Delta_{max}(B_p, B_q)$

Merge (B_p, B_q)

Else

Preserve boundary

End If

End For

Generate $S(B_p, B_q)$ from merged segments

Return

3. RESULTS

The experimental analysis is using a standard dataset from the Middlebury stereo images [27]. These images have a maximum of 1800×1500 pixels, 0-190 disparity range and preprocessing involved normalization process. This evaluation protocol involves 15 training images with multi resolution stereo image pairs with ground-truth and these training images are contained with complex regions. The performance is measured by two key error attributes. The non-occ calculates accuracy only in non-occluded regions visible in both images and reflects baseline stereo matching ability. Then, all errors evaluate accuracy across the entire image with occluded areas, making it a more difficult measure of robustness in handling discontinuities, low-texture, and plain color regions. The experiments were run on a Windows 11 operating system with an Intel Core i5 processor, 16 GB RAM, NVIDIA GeForce RTX 4060 GPU, OpenCV 4.3, Microsoft Visual Studio 2022, C++ and single-threaded execution.

Depth map results from the Adirondack stereo image are illustrated in Figure 2. Problems are illustrated in Image A and Image C where depth surfaces are noisy, and edges appear blurry before boundary segmentation is applied. Common techniques for estimating depth can be challenging. Instances such as the armrest of a chair in Image E demonstrates that small details are often lost due to mismatches and insufficient gradients. Common segmentation techniques do not improve precision for all cases. Noisy depth surfaces and blurry edges are significant issues for both Images A and segmented color Image B. Improved depth maps are achieved through the use of boundary segmentation for Image B and Image D. Without boundary segmentation the depth maps do not perform as well in these experimental results. Moreover, smaller regions of the image are created through the use of edges and gradients. Difficult regions are not isolated effectively which can lead to errors penetrating boundaries. Operating within defined areas results in the depth map developing clearer edges enhancing precision. This observation can be seen in the zoomed section. The proposed boundary segmentation clearly reconstructed the armrest area as depicted in segmented color Image F. Image G shows the depth map recovery result which addresses accurate contour the depth values to achieve a smooth clear depth surface. Gradient consistency sees enhancement through this approach maintaining a true structure which is often not achieved by common available methods. In relation to current available techniques [30],[34],[35], the proposed boundary segmentation preserves complex regions while enhancing local precision. Local accuracy is improved, which is often characterized by an attempt to maintain overall smoothness. Depth maps created have enhanced appearances and exhibit higher accuracy in their shapes. Improvements enhance the utility of depth estimation for applications such as 3D reconstruction, object recognition and scene comprehension. An accurate depth map for stereo matching algorithm has been achieved with boundary segmentation.

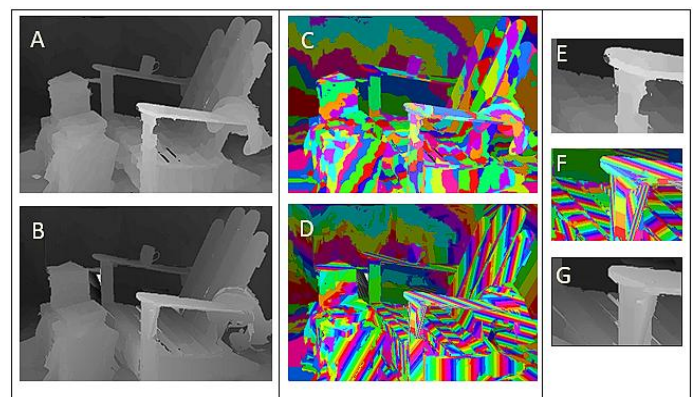


Figure 2. The depth map result of Adirondack image. A) before segmentation, B) the proposed segmentation method, zoomed inaccurate armrest image, C) before segmented process, D) the proposed method after segmented, E) zoomed armrest image before segmented process, F) segmented region of zoomed armrest image and G) smooth and recovered armrest depth map image



Figure 3. Depth map results of PianoL and Pipes images from the proposed method and state-of-the-art methods.

Table 1. SSIM and RMSE evaluation of depth map based on figure 3

Method	SSIM (PianoL)	RMSE (PianoL)	SSIM (Pipes)	RMSE (Pipes)
Proposed	0.92	3.1	0.90	3.5
ACMC	0.88	4.2	0.85	4.8
PPEP-GF	0.86	4.5	0.84	5.0
Dense-CNN	0.90	4.0	0.89	4.0
Esmea	0.87	4.0	0.86	4.6

Table 1 tabulates the SSIM and RMSE results for each image from figure 3. The proposed method achieves the best overall performance, with the highest SSIM (0.92 for PianoL, 0.90 for Pipes) and lowest RMSE (3.1 and 3.5). Dense-CNN, ACMC and Esmea deliver moderate results, with SSIM around 0.85–0.9 and RMSE between 4.0–4.8. PPEP-GF performs weakest, with the lowest SSIM (0.84–0.86) and highest RMSE (4.5–5.0). Overall, the proposed method demonstrates more accurate depth map estimation with good boundary preservation and accuracy.

4. DISCUSSION

Figure 3 shows the depth map results from existing methods in [27] which uses two images known as PianoL and Pipes images for comparison. These images contain of complex regions. Areas lacking texture such as flat surfaces present challenges for stereo matching due to the absence of distinct features. In these regions improved results are achieved with the proposed method. Smooth depth maps are produced minimizing the noise. ACMC and PPEP-GF frequently create inconsistent areas as well as results that lack clarity. There has been some enhancement noted with Dense-CNN yet minor issues persist. Spotted and unstable depth maps are provided by Esmea in these flat regions making its performance the least favorable. Surfaces in solid colors such as the piano lid or pipes (color surfaces) exhibit consistent brightness levels and may lead to inaccurate pairings. Matches can be thrown off by these surfaces. Variations in brightness will not be present on such plain surfaces. Issues with streaks are encountered by ACMC, yet reductions in these streaks are achieved by PPEP-GF, though complete removal is not accomplished. Dense-CNN produces clearer results, but a small degree of banding effect is still noticeable. Depth smoothness is not maintained by Esmea resulting in unreliable and fragmented outcomes. Achieving consistency is frequently lacking in Esmea's outputs. Sharpness is well-maintained by the proposed method effectively. ACMC does not have defined edges nor does it have clear outlines. Huge errors are formed around edges by PPEP-GF. Edges are better detected by Dense-CNN compared to PPEP-GF method, yet some smoothing occurs. Blurred or irregular edges are generated by Esmea resulting in unclear shapes. To sum up, optimal results are obtained with the proposed method for low-texture, plain color and discontinuity regions for both images. Smooth depth is harmonized in flat regions while sharp boundaries are maintained. Both classic approaches (ACMC PPEP-GF) and learning-based techniques (Dense-CNN Esmea) are not as effective. A robust option arises for challenging indoor environments where precise depth matters significantly.

Table 2. Ablation studies of the proposed and state-of-the-art Methods from the Middlebury evaluation system

Method	Non-occ error (%)	All error (%)	Run time (s)	Time Complexity per pixel	D1 (%)	Accuracy (EPE: End-Point-Error)
Proposed	7.59	10.90	0.011	$O(N \cdot d)$	4.1	0.5-1.0
Esmea [30]	8.01	11.30	0.020	$O(N \cdot d)$	4.2	0.53-1.10
Proposed w/o gST	9.24	12.05	0.009	$O(N \cdot d)$	4.1	0.5-1.0
Proposed w/o bSeg	8.95	11.78	0.010	$O(N \cdot d)$	4.1	0.5-1.0
Dense-CNN [36]	4.21	12.10	0.60	$O(N \cdot d^2)$	5.0	1.0-1.5
RDNet [34]	11.30	17.00	0.50	$O(N \cdot d)$	4.8	~1.3
PPEP-GF [35]	9.98	17.10	0.65	$O(N \cdot d \cdot k)$	4.8	~1.0
ACMC [3]	9.50	19.40	0.65	$O(N \cdot d \cdot k)$	4.5	0.9-1.25

Table 2 tabulates the performance comparison between several state-of-the-art algorithms that are Esmea [30], PPEP-GF [35], Dense-CNN [36], ACMC [3] and RDNet [34] evaluated using the Middlebury benchmarking evaluation system [27]. This table has six metrics which are non-occ (i.e., error in region without occlusion), all errors (i.e., error in region with all pixels), runtime, computational complexity (i.e., N : number of pixels in the image, d : disparity range (depth search window), k : kernel or neighborhood size used in filtering or aggregation) and ablation studies. The proposed method attains a non-occ error of 7.59% and all error of 10.90%, demonstrating great performance throughout both metrics. The performance of the proposed method without the gradient-based spanning tree (gST) and boundary segmentation (bSeg) approach produces higher error compared to the complete proposed framework. This demonstrates that the full method can reduce errors and improve results. Dense-CNN produces the lowest non-occ error of 4.21%, but all error is slightly higher at 12.10% implying less robustness in complex regions. Esmea attains 8.01% and 11.30% for non-occ and all errors, respectively. The remaining methods such as RDNet, PPEP-GF, and ACMC demonstrate notably higher error rates, with ACMC making the highest overall error at 19.40%.

This ablation study provides an overall comparison with several state-of-the-art methods under the Middlebury evaluation system. The proposed method achieves the fastest runtime 0.011 seconds and efficient complexity $O(N \cdot d)$. With D1 score of 4.1% and EPE accuracy within 0.5–1.0, it achieves accurate disparity estimation at a low computational cost. Esmea [30] demonstrates comparable efficiency but slightly higher error rates and slightly worse EPE accuracy, suggesting that the proposed design offers a tangible improvement in accuracy without any loss in speed. Dense CNN [36] has the lowest non occ error, which shows its ability in non occluded region. However, this comes at slower run time of 0.60 seconds and quadratic complexity $O(N \cdot d^2)$, making it less practical for real time applications. RDNet [34] has an EPE of ~ 1.3 , which shows that it is not robust, despite being memory efficient. PPEP GF [35] and ACMC [3] have a higher runtime (0.65 seconds) due to guided filtering and multi-scale aggregation, respectively. They are still moderate in accuracy with errors around 9.5-10% and EPE values close to 1.0, but none of them achieves the balance of speed and precision demonstrated by the proposed method. In general, the ablation study confirms the effectiveness of the proposed algorithm in terms of the best balance between the accuracy, the runtime and the computational complexity when compared to the similar methods in terms of efficiency and robustness. Taken together, these results demonstrate the novelty and efficiency of the proposed algorithm and the potential of underwater stereo vision applications requiring accuracy, robustness and computational efficiency.

Table 3. Parameter sensitivity analysis of the proposed method

Parameter	Variation	Effect on Error	Effect on Runtime
Disparity Range	128 $\rightarrow$ 192	-0.3% error	+40% runtime
Refinement Iterations	1 $\rightarrow$ 5	Accuracy plateaus after 3	Runtime increases linearly

Table 3 shows the sensitivity experiments of the proposed algorithm under different hyperparameter settings. Increasing the disparity range from 128 to 192 only slightly improves the accuracy (-0.3 % error) but increases the runtime drastically (+40 %). With each refinement iteration, the accuracy improves. This is true for up to three passes after which the accuracy gains while the run time continues to increase. These findings confirm that the proposed method is robust to parameter changes, with optimal trade-offs achieved at moderate settings.

Table 4. Quantitative robustness validation analysis between the selected state of the arts methods

Condition	Proposed Method Error (%)	Esmea [30]	Dense-CNN [36]
Gaussian noise ($\sigma=0.05$)	12.50	13.20	14.00
Low-light (30% brightness)	11.80	12.60	13.50
Textureless regions	15.00	16.50	17.20

Table 4 tabulates the quantitative robustness validation experiments that were performed to test the robustness of the proposed algorithm under challenging visual conditions such as Gaussian noise, low light environments and texture-less regions. These three controlled degradations were employed to the depth map results at this stage for robustness validation analysis. The metric used is error percentage (%), where lower values indicate better performance and higher robustness. First, Gaussian noise with a standard deviation of $\sigma = 0.05$ was added. This value is selected because it represents a moderate and realistic level of image degradation that is strong enough to challenge the proposed algorithm. Second, low-light conditions were generated by reducing image brightness to 30% of the original intensity, thus representing poor illumination environments. Third, textureless regions were constructed by controlling local texture information through image smoothing, resulting in homogeneous areas with limited discriminative features. The disparity estimation performance under these conditions was then quantified using the percentage error metric shown in table 4. The results show that the proposed method always gets lower error rates than Esmea [30] and Dense CNN [36], which proves its stability in different scenarios.

Under the Gaussian noise with standard deviation 0.05, the error obtained by the proposed method is 12.5%, which is better than Esmea (13.2%) and Dense CNN (14.0%). In low-light environment, when brightness is decreased to 30%, the proposed method results in an error of 11.8% which is still less than Esmea (12.6%) and Dense CNN (13.5%). The proposed method achieves a 15.0% error on texture-less regions, which are difficult for stereo correspondence, compared to 16.5% for Esmea and 17.2% for Dense CNN. These results demonstrate the robustness of the proposed approach under noise, illumination variations and low texture surfaces. Dense CNN performs well in non-occluded regions but its accuracy drops more for adverse conditions, indicating its sensitivity to changes in the environment. Esmea is moderately resilient, but never outperforms the proposed method for any of the tested scenarios. The proposed algorithm shows consistent advantages in all the robustness tests, which demonstrates the effectiveness of the gradient tree cost computation and boundary refinement strategies of the framework for the stability of disparity estimation. Overall, the robustness validation verifies that the proposed method is not only efficient and accurate under typical benchmarks, but also robust to real world degradations, making it a reliable solution for practical stereo vision applications.

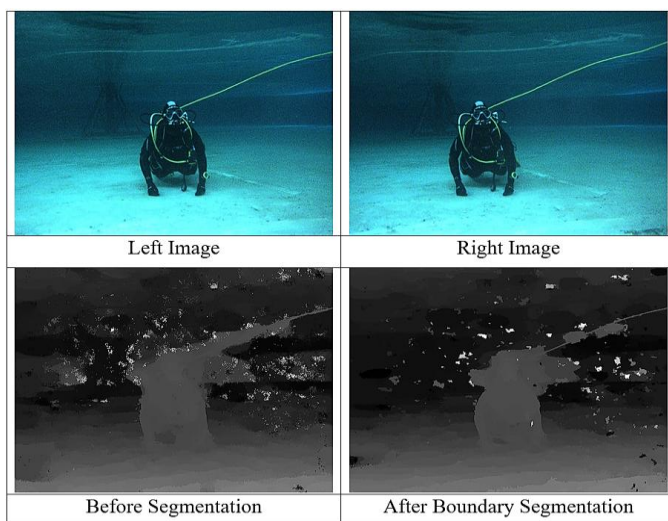


Figure 4. The underwater stereo images from the Caddy Underwater Stereo Dataset [37]

Figure 4 shows the experimental results of underwater stereo images from [37] and the depth map results before and after segmentation process. The depth map reconstruction using gradient-tree structures combined with boundary segmentation demonstrates clear improvements in underwater stereo vision. The preliminary depth output, prior to segmentation, demonstrated blurred contours and significant noise around low-texture regions such as the sandy seabed. These artifacts reduced the reliability of depth estimation and made object boundaries particularly the diver difficult to distinguish. Applying boundary segmentation improved clarity by sharpening object contours and reducing background noise. The diver's outline became more distinct, and structural details such as the hose and seabed were better separated from the

surrounding water. This refinement enhanced the accuracy of depth estimation in complex underwater regions. This improvement highlights the strength of gradient-tree structures in managing intensity variations while boundary segmentation compensates for weak texture regions, which are common in underwater imagery due to scattering and absorption effects.

5. CONCLUSIONS

The proposed work in this article produces encouraging progress in depth map reconstruction for underwater robot vision which the combination of gradient spanning trees and boundary-based segmentation enhances accuracy and reliability. Several experiments on the Middlebury evaluation benchmark validate that the proposed method obtains a good balance among accuracy, computation efficiency and robustness. The proposed algorithm performs better than the comparable methods like Esmea, PPEP-GF, Dense-CNN, ACMC and RDNet with the non-occ error of 7.59%, all error of 10.90% and EPE accuracy in the range of 0.5-1.0 pixels, while keeping competitive performance against deep learning-based Dense CNN in non-occluded regions. Importantly, the proposed work achieves the fastest runtime (0.011 s) with linear complexity $O(N \cdot d)$. Ablation studies highlighted the critical role of gradient spanning tree and boundary refinement in reducing disparity errors, particularly in textureless and occluded regions. Sensitivity analysis further confirmed the stability of the proposed work across parameter variations, with optimal balance achieved at moderate disparity ranges and refinement iterations. Robustness validation under Gaussian noise, low-light conditions, and textureless surfaces demonstrated that the proposed method constantly provides lower error rates compared to established methods. Overall, the results prove that the proposed work offers the potential solution for depth map reconstruction in stereo vision application such as robotics, autonomous navigation and underwater imaging, delivering pixel-level accuracy with low computational overhead. Future work will investigate the extension of the framework to multi view reconstruction as well as the inclusion of adaptive learning strategies to further improve the generalisation across diverse environments.

Author Contributions: Conceptualization, Rostam Affendi Hamzah; methodology, Nurliyana Abd Mutalib; software, Ken Prameswari Caesarella Aryaputri; validation, Ahmad Fauzan Kadmin; writing—original draft preparation, Masrullizam Mat Ibrahim, Adi Irwan Herman.

Funding: This research was funded by the Universiti Teknikal Malaysia Melaka.

Acknowledgments: The authors would like to express their sincere gratitude to Universiti Teknikal Malaysia Melaka (UTeM) for providing the fund.

Conflicts of Interest: “The authors declare no conflict of interest”

REFERENCES

- [1] G. Lu, P. Zhang, B. Song, “Stereo Matching Algorithm Combining Image Segmentation and Improved Belief Propagation”, *Computer Fraud and Security*, vol. 12, pp. 12–23, 2024, doi: 10.52710/CFS.273.

- [2] T. Li, Z. Huang, L. He, J. Wang, "Stereo Matching Algorithm Based on Improved Census Transform and Adaptive Window Optimization", *International Conference on Computer Science and Blockchain*, 2024, pp. 219–224, doi: 10.1109/CCSB63463.2024.10735587.
- [3] M. Zahari, R.A. Hamzah, "Stereo matching algorithm for autonomous vehicle navigation using integrated matching cost and non-local aggregation", *Bulletin of Electrical Engineering and Informatics*, vol. 12, no 1, 2023, pp. 328–337, doi: 10.11591/EEI.V12I1.4122.
- [4] C. Lv, J. Li, S. Tang, "Stereo Matching Algorithm Based on HSV Color Space and Improved Census", *Transform. Math Probl Eng*, vol. 2021, no. 1, 2021, 1857327, doi: 10.1155/2021/1857327.
- [5] Y. Hou, C. Liu, Y. Liu, "Stereo matching algorithm based on improved Census transform and texture filtering", *Optik*, vol. 249, 2021, p. 168186, doi: 10.1016/J.IJLEO.2021.168186.
- [6] A. F. Kadmin, R. A. Hamzah, T. F. T. Wook, "Local stereo matching algorithm using modified dynamic cost computation", *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, 2021, pp. 1312–1319, doi: 10.11591/IJEECS.V22.I3.PP1312-1319.
- [7] Li Z., Huang J., Wang W., & Huang Y. (2024). A new stereo matching algorithm based on improved four-moded census transform and adaptive cross pyramid model. *Electronic Research Archive*, vol. 32, no. 7, 4340–4364. doi: 10.3934/era.2024195.
- [8] D. Wei, Y. Zhang, C. Li, "Robust line segment matching via reweighted random walks on the homography graph," *Pattern Recognit*, vol. 111, 2021, p. 107693, doi: 10.1016/J.PATCOG.2020.107693.
- [9] H. Liu, "Stereo matching algorithm based on two-phase adaptive optimization of AD-census and gradient fusion", *IEEE International Conference on Real-Time Computing and Robotics*, 2021, pp. 726–731, doi: 10.1109/RCAR52367.2021.9517511.
- [10] D. Xi, H. Yang, B. Tan, "Stereo matching algorithm based on improved census transform and minimum spanning tree cost aggregation", *J Vis Commun Image Represent*, vol. 98, 2023, pp. 104023–104038, doi: 10.1016/J.JVCIR.2023.104023.
- [11] S. Yang, X. Lei, Z. Liu, G. Sui, "An efficient local stereo matching method based on an adaptive exponentially weighted moving average filter in SLIC space", *IET Image Process*, vol. 15, no. 8, 2021, pp. 1722–1732, doi: 10.1049/IPR2.12140.
- [12] B. Hou, H. Ma, Z. Li, W. Zhang, "Segment-tree based stereo matching algorithm of segmented regions collaborative optimization", *International Conference on Electronic Information Engineering*, 2022, pp. 209–216, doi: 10.1117/12.2634688.
- [13] Z. Huang, G. Wan, X. Tao, "Adaptive Local Stereo Matching Based on Improved Census Transform", *Proceedings of China Automation Congress*, 2023, pp. 2144–2149, doi: 10.1109/CAC59555.2023.10451091.
- [14] A. F. Kadmin, R. A. Hamzah, M. N. Abd Manap, S. F. Abd Gani, "Improved stereo matching algorithm based on census transform and dynamic histogram cost computation", *International Journal of Emerging Technology and Advanced Engineering*, vol. 11, no. 8, 2021, pp 48–57, doi: 10.46338/IJETA0821_07.
- [15] L. Zhao, F. Guo, B. Zhou, "A Generalized Voronoi Diagram-Based Segment-Point Cyclic Line Segment Matching Method for Stereo Satellite Images", *Remote Sensing*, vol. 16, no. 23, 2024, p. 4395, doi: 10.3390/RS16234395.
- [16] J. Zhang, Y. Zhang, C. Wang, C. Qin, "Binocular stereo matching algorithm based on MST cost aggregation", *Mathematical Biosciences and Engineering*, vol. 18, no. 4, 2021, pp. 3215–3226, doi: 10.3934/MBE.2021160.
- [17] X. Li, Y. Bai, "Stereo Matching Algorithm Based on Improved Census Transform and Fusion Cost", *International Conference on Communication Technology and Information Technology*, 2023, pp. 176–180, doi: 10.1109/ICCTIT60726.2023.10435961.
- [18] H. Wei, L. Meng, "An accurate stereo matching method based on color segments and edges," *Pattern Recognit*, vol. 133, 2023, p. 108996, doi: 10.1016/J.PATCOG.2022.108996.
- [19] C. Braunstein, E. Ilg, V. Golyanik, "Quantum-Hybrid Stereo Matching with Nonlinear Regularization and Spatial Pyramids", *International Conference on 3D Vision*, 2024, pp. 1340–1349, doi: 10.1109/3DV62453.2024.00121.
- [20] X. Chen, W. Zhang, Y. Hou, L. Yang, "Improved stereo matching algorithm based on multi-scale fusion", *Journal of Northwestern Polytechnical University*, vol. 39, no. 4, 2021, pp. 876–882, doi: 10.1051/JNWPU/20213940876.
- [21] G. Yao, "Quasi-Dense Matching for Oblique Stereo Images through Semantic Segmentation and Local Feature Enhancement", *Remote Sensing*, vol. 16, no. 4, 2024, p. 632, doi: 10.3390/RS16040632.
- [22] S. Wu, X. Wang, J. He, X. Song, "Research on Trans. Line Stereo Matching Based on Twin Residual Network", *Conf. on Electronics and Electrical Eng. Tech.*, 2022, pp. 90–97, doi: 10.1109/EEET58130.2022.00024.
- [23] K. Yit Kok, P. Rajendran, "A local multi-block with various constraints approach for stereo vision in StereoPi", *Measurement*, vol. 200, 2022, p. 111617, doi: 10.1016/J.MEASUREMENT.2022.111617.
- [24] N. Liu, N. Zhao, X. Ouyang, "Context Geometry Volume and Warping Refinement for Real-Time Stereo Matching", *Electronics*, vol. 14, no. 5, 2025, p. 892, doi: 10.3390/ELECTRONICS14050892.
- [25] Z. Xu, J. Wang, Y. Wang, "A Dual Branch Multiscale YStereo Matching Network for High-Resolution Satellite Remote Sensing Images", *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 1v, 2025, pp. 949–964, doi: 10.1109/JSTARS.2024.3502842.
- [26] L. Dong, Y. Han, M. Hu, Y. Wang, "Stereo matching method based on high-precision occlusion-recovering and discontinuity-preserving", *Displays*, vol. 80, 2023, p. 102573, doi: 10.1016/J.DISPLA.2023.102573.
- [27] D. Scharstein, et al., "High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth", *Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 2014, pp. 31–42, doi: 10.1007/978-3-319-11752-2_3/COVER.
- [28] L. Kong, J. Zhu, S. Ying, "Stereo Matching Based on Guidance Image and Adaptive Support Region", *Acta Optica Sinica*, vol. 40, no. 9, 2020, p. 0915001, doi: 10.3788/AOS202040.0915001.
- [29] H. Yaoyu, Z. Weikun, S. Sebastian, "Deep-Learning Assisted High-Resolution Binocular Stereo Depth Reconstruction", *IEEE International Conference on Robotics and Automation*, 2020, pp. 8637–8643.
- [30] S. Fan, W. Sun, J. Zheng, W. Wu, "Accurate edge-preserving stereo matching by enhancing anisotropy", *Signal Process Image Commun*, vol. 114, 2023, p. 116945, doi: 10.1016/J.IMAGE.2023.116945.
- [31] Y. Fu, K. Lai, W. Chen, Y. Xiang, "A pixel pair-based encoding pattern for stereo matching via an adaptively weighted cost", *IET Image Process*, vol. 15, no. 4, 2021, pp. 908–917, doi: doi.org/10.1049/ipr2.12071.
- [32] Y. Zhang, Q. Shen, P. Song, "Local Side Window Algorithm with Tree Segmentation for Stereo Matching", *Laser & Optoelectronics Progress*, vol. 61, no. 22, 2021, pp. 2237007–2237007, doi: 10.3788/LOP240703.
- [33] B. Ruf, J. Mohrs, M. Weinmann, J. Beyerer, "ReS2tAC—UAV-Borne Real-Time SGM Stereo Optimized for Embedded ARM and CUDA Devices", *Sensors*, Vol. 21, No. 11, 2021, p. 3938, doi: 10.3390/S21113938.
- [34] J. Min, J. Kim, C. H Min, M. Choi, "DepthFocus: Controllable Depth Estimation for See-Through Scenes", *arXiv preprint arXiv:2511.16993*, 2025, p. 1, doi.org/10.48550/arXiv:2511.16993.
- [35] Y. Wang, K. Li, L. Wang, J. Hu, D. O. Wu, and Y. Guo, "Adstereo: Efficient stereo matching with adaptive downsampling and disparity alignment", *IEEE Transactions on Image Processing*, vol. 34, 2025, pp. 1204–1218, doi.org/10.1109/TIP.2025.3540282.
- [36] Z. Congxuan, W. Junjie, J. Shaofeng, "Dense-CNN: Dense convolutional neural network for stereo matching using multiscale feature connection", *Signal Processing: Image Communication*, vol. 95, 2021, p. 116285, doi.org/10.1016/j.image.2021.116285.
- [37] A. Gomez Chavez, A. Ranieri, D. Chiarella and A. Birk, "Underwater Vision-Based Gesture Recognition: A Robustness Validation for Safe

Human–Robot Interaction”, IEEE Robotics & Automation Magazine, vol. 28, no. 3, pp. 67-78. doi.org/10.1109/MRA.2021.3075560

- [38] G., Li, S. Huang, Z. Yin, J. Li, and K. Zhang, “Underwater Refractive Stereo Vision Measurement and Simulation Imaging Model Based on Optical Path”, Journal of Marine Science and Engineering, vol. 12, no. 11, 2024, p. 1955. doi.org/10.3390/jmse12111955.
- [39] L. Zhu, Y. Gao, J. Zhang, Y. Li, and X. Li, “Reliable and Effective Stereo Matching for Underwater Scenes. Remote Sensing”, vol. 16, no. 23, 2024, p. 4570. doi.org/10.3390/rs16234570
- [40] Q. Lv, J. Dong, Y. Li, S. Chen, and W. Wang, “UWStereo: A Large Synthetic Dataset for Underwater Stereo Matching”, arXiv preprint, arXiv:2409.01782. 2025. doi.org/10.48550/arxiv.2409.01782.



© 2026 by the Nurliyana Abd Mutalib, Rostam Affendi Hamzah, Ken Prameswari Caesarella Aryaputri, Ahmad Fauzan Kadmin, Masrullizam

Mat Ibrahim, and Adi Irwan Herman. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).