Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 13, No. 4, December 2025, pp. 1050∼1064
ISSN: 2089-3272, DOI: 10.52549/ijeei.v13i4.6686

Robust Stereo Matching for Driver Assistance Systems Under Adverse Driving Conditions

Vinh Dinh Nguyen[1], Nhan Huu Tran[1]
[1]Department of Information Technology, FPT Univeristy, Can Tho Campus, Can Tho, Ninh Kieu, Vietnam

Article Info

Article history:

Received May 20, 2025
Revised Dec 5, 2025
Accepted Dec 18, 2025

Keywords:

complex driving condition
adverse weather conditions
depth estimation
deep learning
stereo matching

ABSTRACT

Deep stereo networks perform effectively when both training and testing data come from the same domain. However, their accuracy tends to drop significantly in efficiency-focused target scenarios due to domain shifts between training and testing datasets. These shifts often arise from differences in factors such as color, lighting, contrast, and texture. Additionally, the architecture of deep networks generally results in processing times that are unsuitable for real-time applications. To address these issues, this paper proposes a lightweight and robust stereo matching approach tailored for diverse driving environments. It leverages attention mechanisms for feature extraction and uses evolutionary algorithms for optimizing parameters. The method outperforms existing deep learning and traditional stereo matching techniques in terms of both processing speed and the percentage of bad pixels, as demonstrated on three challenging outdoor datasets: KITTI, HCI, and Driving Stereo. These results indicate that the proposed solution is highly effective for real-world applications where both precision and flexibility are essential.

Copyright © 2025 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Vinh Dinh Nguyen
Department of Information Technology, FPT University
No.600, Nguyen Van Cu St. (extend), An Binh Ward, Ninh Kieu Dist., Can Tho, Vietnam
Email: vinhnd29@fe.edu.vn

1. INTRODUCTION

Advanced driver-assistance systems enhance safety by reducing accident risks through critical traffic and road data collection, such as distance, obstacle identification, traffic signs, signals, and nearby vehicles. Key features like distance measurement and 3D obstacle detection, often achieved with stereo vision, are essential for effective traffic scenario analysis [1, 2]. Stereo vision algorithms can be divided into two categories: non-deep learning-based and deep learning-based methods. Non-deep learning methods, like pSGM, iSGM, and SGM offer fast processing; however, their performance often degrades in challenging environments such as textureless or specular regions. Since then, the field has advanced significantly, driven by new techniques and paradigms inspired by breakthroughs in other deep learning areas. Transformer-based and iterative refinement architectures are key examples of the field's evolution, showcasing the potential for continued improvements in accuracy and efficiency [2]. However, these approaches typically introduce higher computational and memory costs, making them less practical for real-time ADAS applications.

Recent advancements in stereo matching have introduced a variety of methods to enhance disparity estimation accuracy and computational efficiency. Li et al. [3] proposed a cascaded recurrent network that effectively captures contextual information but increases computational complexity. Zhong et al. [4] addressed underwater distortion through unsupervised adaptation, though performance suffers in turbid environments. Dai et al. [5] proposed real-time disparity prediction, efficient yet less robust in complex scenes. Lai et al.

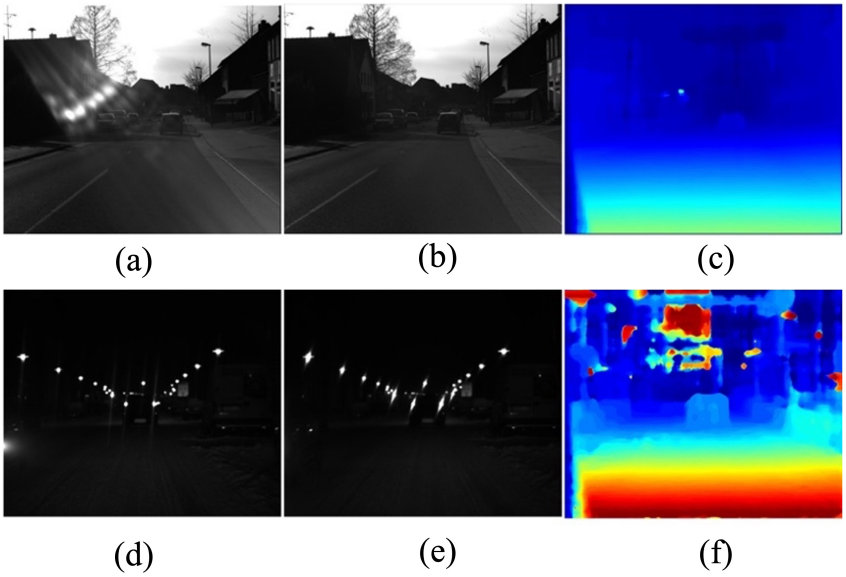


Figure 1. Experimental results of the proposed method under the HCI dataset. (a), (b), and (c) represent the left image, right image, and the disparity of the proposed method, respectively, under sun flare conditions. (d), (e), and (f) represent the left image, right image, and the disparity results of the proposed method, under night and snow conditions.

[6] improved textureless area matching using adaptive weights but with added computation. Xu et al. [7] and Wang et al. [8] enhanced accuracy through iterative refinement and frequency selection, though both methods are computationally intensive. Li et al. [9] employed local structure guidance, which may struggle with structural variance. Cheng et al. [10] achieved high FPS for XR but compromised on depth accuracy. Yue et al. [11] used semi-supervised learning to manage imperfect data, with results tied to label quality. Kwon and Zell [12] introduced image-coupled propagation improving accuracy but increasing processing load. Song et al. [13] applied Bayesian methods for surgical imagery, effective yet computationally demanding. Quan et al. [14] improved specular region accuracy via multi-spectral attention, though data-hungry. Zhang et al. [15] fused temporal-spatial cues for dynamic scenes, at the cost of processing speed. Multi-scale attention and contextual refinement by Zhang et al. [16] and Liu et al. [17] yielded precise results but demanded more resources. Wang et al. [18] and Zhang et al. [19] demonstrated that feature fusion and attention mechanisms enhance accuracy, though training and inference remain resource-intensive. Chen et al. [20] incorporated graph networks for global context, facing scalability issues. Adaptive cost aggregation [21] and transformer networks [22] improved matching in complex regions, yet with greater computational costs. Deng et al. [23] fused multi-scale information to enhance matching accuracy, though at the expense of increased computational complexity. Qi et al. [24] introduced an adaptive support window, improving accuracy in weak-textured regions but with added computation. Zhang et al. [25] proposed LiDAR-guided stereo matching, enhancing robustness but requiring additional sensor data. Yang et al. [26] optimized hierarchical graph structures for urban 3D reconstruction, though scalability remains a concern. Iqbal et al. [27] guided deep learning with expert knowledge, improving depth estimation but with increased model complexity. Tosi et al. [28] provided a comprehensive survey on deep stereo matching, offering insights into various approaches but lacking new algorithmic contributions. Yang et al. [29] proposed a non-local cost aggregation method that improves matching accuracy but increases computational complexity. Zhang et al. [30] introduced a guided aggregation network that enhances disparity estimation but requires additional training data. Gu et al. [31] developed a cascade cost volume for high-resolution stereo matching, improving accuracy but at the cost of increased processing time. Pang et al. [32] applied residual learning for stereo matching, achieving better results but with higher computational demands. Zhang et al. [33] utilized multi-dimensional residual dense attention networks for stereo matching, enhancing performance but requiring more computational resources. Zhang et al. [34] employed long-range attention networks for multi-view stereo, improving depth estimation but with increased computational complexity. Zhou et al. [35] introduced DeepTAM for deep tracking and mapping, achieving accurate results but with higher

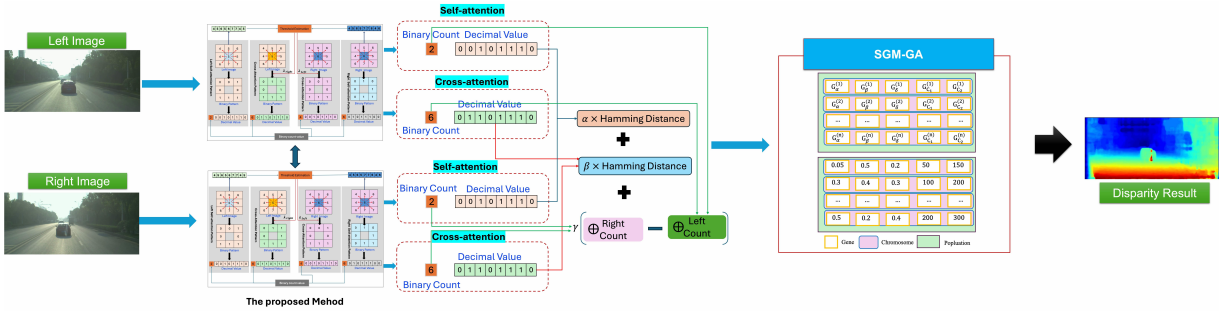


Figure 2. The workflow of the proposed system.

computational requirements. Yildirim [36] studied cybersickness during VR gaming, providing insights into user experience but not directly related to stereo matching. Kang et al. [37] investigated the effect of outside view on attentiveness in using see-through type augmented reality devices, offering valuable information for AR applications but not directly applicable to stereo matching. Although the above techniques provide specific improvements, they still face a common limitation: a balance between robustness, accuracy, and real-time speed has not yet been achieved.

Furthermore, performance degradation becomes evident when deploying deep models in diverse real-world environments, particularly under illumination changes, weather variations, and domain shifts between synthetic and real datasets. These challenges highlight the need for a stereo matching system that is both robust and computationally efficient for deployment in autonomous driving.

Fig. 1 shows results from the HCI dataset [38], where the performance of existing deep learning-based method was decreased under difficult cases in real-world conditions, while our proposed system still obtained the stable performance. This study presents several novel contributions:

- A robust local feature extraction method using the attention mechanism for various driving conditions.
- An efficient stereo data cost function that combines the benefits of census transform cost, local attention cost, and binary count cost.
- An effective evolutionary algorithm for optimizing the hyperparameters of the proposed methods.
- A real-time, robust stereo matching system designed for diverse road conditions, optimized for GPU performance.

The rest of this paper is organized as follows: Section II introduces and explains the proposed algorithm. Section III presents the results and discusses the findings of the proposed method. Finally, Section IV concludes the paper.

2. THE PROPOSED METHOD

This section outlines the proposed method to enhance stereo matching performance in challenging driving conditions. First, the local self and cross-attention feature (SACA) is introduced, followed by the proposed SACA-based strategy for computing stereo data cost. The computed cost is integrated with SGM [39] for disparity map generation. Finally, the proposed evolutionary algorithms are employed to optimize hyperparameters. The overall workflow is illustrated in Fig. 2. SACA is designed to integrate both self-attention and cross-attention in local regions to effectively encode the contextual relationships within each image as well as the correspondence between stereo pairs. Unlike global attention models that require large computational resources, SACA focuses only on compact neighborhoods, ensuring high robustness while maintaining low computational cost for real-time stereo matching.

The Stereo Adaptive Cost Aggregation (SACA) method adaptively aggregates matching costs by adjusting the support region based on local structural information. Instead of relying on a fixed or predefined cost window, SACA dynamically modifies the aggregation scope according to texture strength and edge boundaries. This adaptive mechanism helps preserve disparity discontinuities and improves matching robustness in weakly

textured and occluded areas. As a result, SACA achieves more reliable cost estimation compared to traditional local aggregation approaches that often blur object boundaries.

2.1. Self and Cross Attention Feature: SACA

Inspired by the attention mechanism in the Transformer architecture [22], self-attention links different positions within a sequence to create its representation, and has been successful in tasks like reading comprehension and summarization. Cross-attention enhances information exchange between the encoder and decoder. This research explores using a local self-cross-attention feature to improve stereo matching performance. Unlike Transformer-based methods [22] that equire large networks and multiple GPUs for disparity generation and refinement, our approach integrates attention into local feature computation, making it efficient and not reliant on heavy training or powerful GPUs for inference.

2.1.1. Local Attention Feature-LAF

From the perspective of stereo matching, we can consider local attention as a mechanism to identify and extract stable features from local image regions that are not significantly affected by noise. This idea builds on previous research on extracting stable features from regions of interest (ROI) [40]. Among the existing approaches, we found that the Census transform is a dominant solution due to its computational simplicity and robustness to illumination changes, as it focuses on the relative ordering of pixels rather than their intensity values. Therefore, in this research, we use the Census transform [40] as a method to extract local attention features LAF, as follows:

$$\begin{aligned} &LAF\left(Z_c\right) = \bigotimes_{n=1}^{N} \chi\left(Z_c, Z_n, \psi\right) \\ &\chi\left(Z_c, Z_n, \psi\right) = \begin{cases} 1 \; if Z_c - Z_c > \psi \\ 0 \; if Z_c - Z_c \leqslant \psi \end{cases} \end{aligned} \tag{1}$$

where Z_c and Z_n are the location of center pixel c and the neighbor pixels n, respectively. However, the original Census transform is not sufficient under extremely noisy conditions. Therefore, we investigate a maximum density approach in equation (1) to improve the performance of the Census transform by introducing a parameter ψ based on previous research [41] as follows:

$$\begin{aligned} &\psi = I\left(Z_{\omega *}\right) \\ &Z_{\omega *} = argmax_{Z_i, i \in N} \left[\sum_{n=1}^{N} 1_{I(Z_i)=I(Z_n)}\right] \end{aligned} \tag{2}$$

where $I(Z_n)$ is the pixel intensity value at the local Z_n of the left (or right) image. Thus, our proposed LAF can be considered a local attention approach because it encodes the relationship between the center pixel and its neighboring pixels, as well as the pixel value with the maximum density in the local region.

2.1.2. Local Cross Attention Feature

Most existing local pattern methods for stereo matching, such as Census, Local Ternary Census, Generalized Census Transform, and local density encoding [20], were designed to handle information only from the local left or right image regions. They do not consider the relationship between the left and corresponding right image regions for encoding the features simultaneously. Therefore, this research aims to introduce an efficient method to encode the features of the local regions by considering information from both the left and right image regions simultaneously. This research introduces a method to compute the local features for both the left $LCAF_{Left}$ and right image regions $LCAF_{Right}$ by considering their relationships as follows:

$$\begin{aligned} &LCAF_{Left}\left(Z_c\right) = \bigotimes_{n=1}^{N} \chi\left(Z_c, Z_n, \lambda_{Right}\right) \\ &LCAF_{Right}\left(Z_c\right) = \bigotimes_{n=1}^{N} \chi\left(Z_c, Z_n, \lambda_{Left}\right) \end{aligned} \tag{3}$$

where the λ_{Left} and λ_{Right} are parameters to present how much information are exchange to right local pattern computation and left pattern computation, respectively, as follows:

$$\begin{aligned} &\lambda_{Left} = I_{Right}\left(Z_{\omega_{Left}}\right) \\ &Z_{\omega_{Left}} = argmax_{Z_i, i \in N} \left[\sum_{n=1}^{N} 1_{I_{Right}(Z_i)=I_{Right}(Z_n)}\right] \\ &\lambda_{Right} = I_{Left}\left(Z_{\omega_{Right}}\right) \\ &Z_{\omega_{Right}} = argmax_{Z_i, i \in N} \left[\sum_{n=1}^{N} 1_{I_{Left}(Z_i)=I_{Left}(Z_n)}\right] \end{aligned} \tag{4}$$

where $I_{Left}(Z_n)$ and $I_{Right}(Z_n)$ are the pixel intensity value at the local Z_n of the left and right images, respectively.

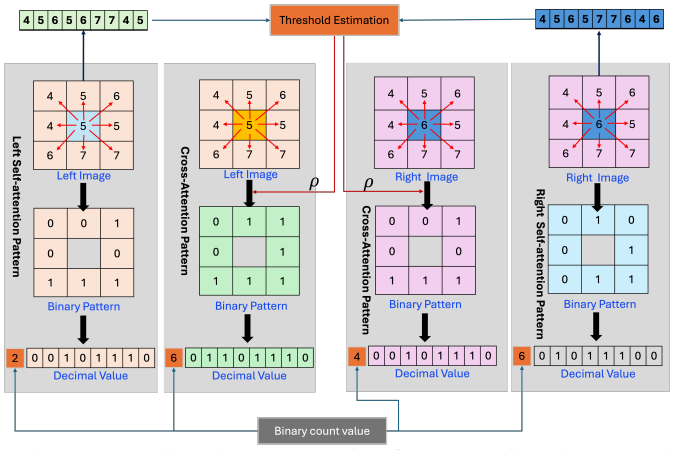


Figure 3. The proposed local cross-attention for computing the stereo data cost.

2.2. Stereo Matching Data Cost Based on SACA

This section describes how we leverage the benefits of SACA (Fig. 3) to develop a robust stereo matching cost. The proposed matching cost includes the data cost from the local attention feature (LAF) and the data cost from the local cross attention feature (LCAF). Therefore, this research introduces an efficient approach to combine the benefits of these two matching costs as follows:

$$\begin{aligned} &D_{SACA}\left(p, f_p\right) = \\ &\alpha \times d_H\left[LAF_{left}\left(x, y\right), LAF_{right}\left(x - f_p, y\right)\right] + \\ &\beta \times d_H\left[LCAF_{left}\left(x, y\right), LCAF_{right}\left(x - f_p, y\right)\right] \\ &d_H\left[q, k\right] = \sum_{i=1}^{N-1} 1_{\{q_i \neq k_i\}} \end{aligned} \tag{5}$$

where $p = (x, y)$ represents the location of pixel p in the image, d_H is the Hamming distance, α is a parameter designed to determine how much the LAF feature contributes to the final cost value, and β is a parameter designed to determine how much the $LCAF$ feature contributes to the final cost value. To further enhance the effectiveness of the proposed matching cost under difficult conditions, this research continues to integrate the binary count feature, which can detect relative pixel differences, helping to maintain accurate matching even in difficult areas, such as low contrast or homogeneous textures, into the proposed system as follows:

$$\begin{aligned} &D_{final}\left(p, f_p\right) = D_{SACA}\left(p, f_p\right) + \delta \times D_{binary_count}\left(p, f_p\right) \\ &D_{binary_count}(x, y) = \\ &\left[\sum_{i=1}^{N} LAF_{left}(i) + \sum_{i=1}^{N} LCAF_{left}(i)\right] - \\ &\left[\sum_{i=1}^{N} LAF_{Right}(i) + \sum_{i=1}^{N} LCAF_{Right}(i)\right] \end{aligned} \tag{6}$$

where binary count information was calculated by considering the local and cross attention mechanism between left and right image paths. Thus, the energy function, $F_u\left(p, f_p\right)$, that assigns disparity, f_p, to pixel p along direction u, as defined in [40], is as follows:

$$\begin{aligned} F_u(p, f_p) = D_{final}(p, f_p) + \min \left\{ \begin{array}{l} F_u(p-u, f_p), \\ F_u(p-u, f_p - 1) + C_1, \\ F_u(p-u, f_p + 1) + C_1, \\ \min_i E_u(p-u, i) + C_2 \end{array} \right\} \\ - \min_k F_u(p-u, k) \end{aligned} \tag{7}$$

where the penalties C_1 and C_2 are designed to estimate disparity changes and disparity discontinuities, respectively, and are pre-defined during the operation of the SGM. However, accurately estimating these two

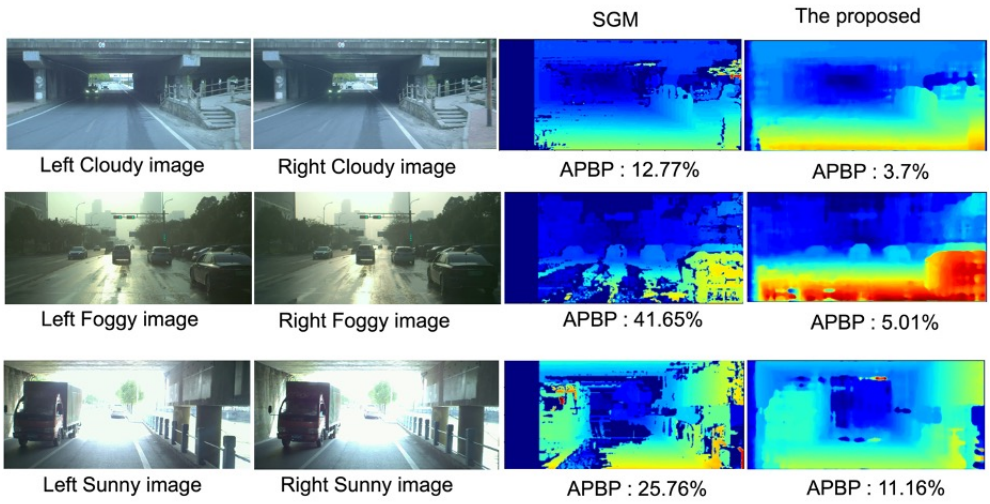


Figure 4. The experimental results on the DrivingStereo dataset are presented. The first, second, and third rows show the results of the SGM and the proposed method under cloudy, foggy, and sunny conditions, respectively.

penalties is challenging. Therefore, this research proposes a method to dynamically estimate C_1 and C_2 using the proposed Genetic Algorithm, which is introduced in the following sections.

2.3. The proposed GA

It is challenging to estimate the optimal thresholds for α, β, δ, C_1, and C_2. Therefore, this research introduces a method to find the best values for these hyperparameters using a genetic algorithm (GA). Initially, a chromosome structure was defined to represent the possible values of α, β, δ, C_1, and C_2. Next, a fitness function was developed to evaluate the fitness of each chromosome. Tournament selection and crossover were applied to select and generate stronger chromosomes for subsequent generations. Finally, after meeting the termination criteria, the strongest surviving chromosomes were identified as the final optimal parameters.

2.3.1. Chromosome Structure

Choosing a suitable genotype representation is crucial for minimizing the processing time in genetic algorithms. In genetic algorithms (GAs) and genetic programming, three commonly used genotype representations are string, matrix, and tree structures [42]. This paper introduces a chromosome structure using string representation due to its flexibility and ease of extending the chromosome size. Each chromosome is organized into five genes: $G_{\alpha}^{(n)}$, $G_{\beta}^{(n)}$, $G_{\delta}^{(n)}$, $G_{C_1}^{(n)}$, and $G_{C_2}^{(n)}$, which represent the possible values of α, β, δ, C_1, and C_2, respectively.

2.3.2. Initial Population:

The population of chromosomes, representing the potential values of each gene, is initialized randomly according to the following constraints. $G_{\alpha}^{(n)}$, $G_{\beta}^{(n)}$, $G_{\delta}^{(n)}$, and G_{δ} are randomly generated within the range [0, 1], while C_1, and C_2 are randomly generated within the range [0, 1000].

2.3.3. Fitness function

The fitness function $fitness_{dis}^{\alpha,\beta,\delta,C_1,C_2}$ was designed to evaluate the accuracy of the disparity results after using the given chromosome parameters including $G_{\alpha}^{(n)}$, $G_{\beta}^{(n)}$, $G_{\delta}^{(n)}$, $G_{C_1}^{(n)}$ and $G_{C_2}^{(n)}$ as follows:

$$fitness_{dis}^{\alpha,\beta,\delta,C_1,C_2} = \sqrt{(\underset{f_p}{argmin} F_u^{\alpha,\beta,\delta,C_1,C_2}(p, f_p) - GT)^2} \tag{8}$$

Where GT represents the disparity ground truth. Once the fitness value for each chromosome is calculated, a pairwise tournament selection is performed to determine the top chromosomes for reproduction in the next generation. After the selection, linear crossover and mutation are applied to generate new offspring for the following generation.

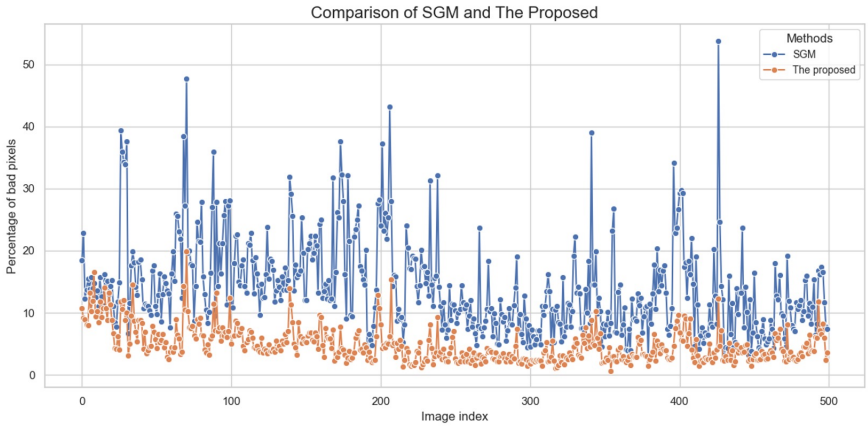


Figure 5. The experimental results of the SGM-based method and the proposed method using 500 stereo image under cloudy from DrivingStereo dataset.

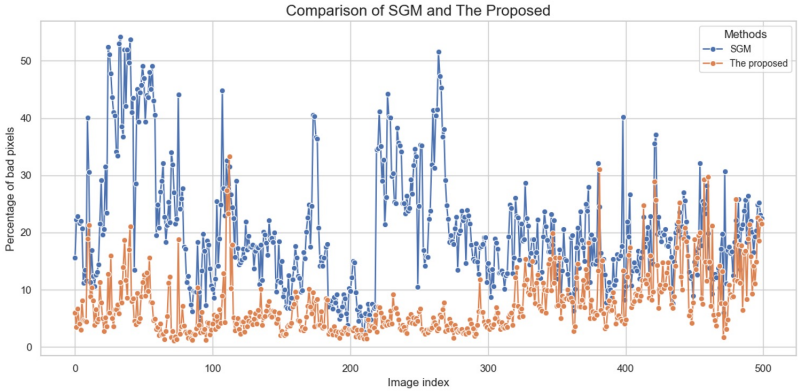


Figure 6. The experimental results of the SGM-based method and the proposed method using 500 stereo image under Foggy conditions from DrivingStereo dataset.

3. RESULTS AND DISCUSSION

3.1. Dataset, Compared Methods, and Parameter Settings

To evaluate the performance of the proposed algorithm, three popular stereo datasets were used: KITTI 2012 and 2015 [43], HCI [38], and DrivingStereo [44]. The KITTI dataset provides stereo images with available ground-truth disparity in urban, rural, and highway environments, featuring dynamic objects and challenging lighting conditions (e.g., reflective regions). The HCI dataset includes stereo images without ground-truth information, captured under more challenging driving conditions, such as blinking arrows, car-truck interactions, crossing cars, flying snow, night driving, snow, rain blur, rain flares, reflecting cars, shadows on trucks, sunflare, and wet autobahn scenarios. DrivingStereo is a large-scale stereo vision dataset with available disparity ground truth, specifically designed for autonomous driving research. It contains over 180,000 stereo image pairs captured in diverse real-world driving scenarios, including varying weather, lighting, and seasonal conditions, such as rain, fog, snow, and nighttime driving. The proposed method was not designed based on deep learning architecture. However, in order to provide a comprehensive evaluation of the performance of the proposed method in comparison with other methods, two groups of stereo methods are set up for comparison as follows. Non-deep learning based methods: the proposed-based SACA, SGM (2022) [39], m-SGM (2023) [45], and s-SGM (2024) [46]. Deep learning based methods: FEDL [47] (2021), permutation (2022) [48], and CRD-Fusion (2022) [49], and RAPIDFlow (2024) [50]. The proposed method differs significantly from [47]. In our approach, an attention-based mechanism is introduced to extract more robust and discriminative features, whereas [47] relies on traditional local feature descriptors such as Census. Moreover, we incorporate a Genetic Algorithm (GA) to adaptively optimize the SGM parameters, while [47] adopts fixed and constant SGM settings by designing a predefined function to estimate.

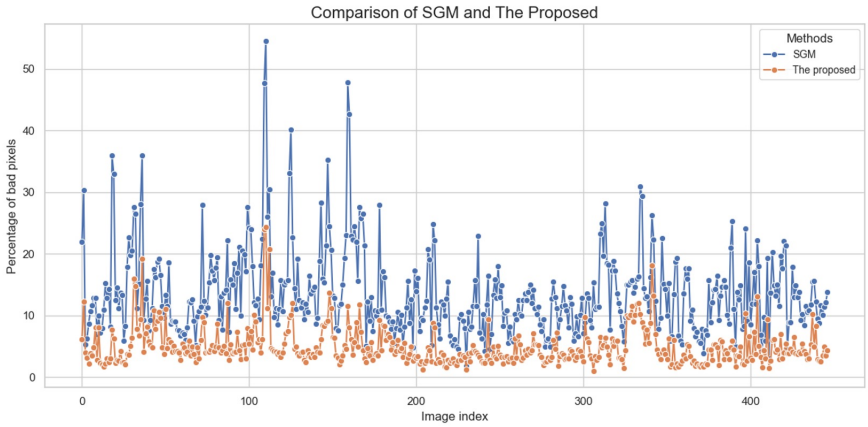


Figure 7. The experimental results of the SGM-based method and the proposed method using 500 stereo image under Sunny from DrivingStereo dataset.

3.2. Evaluation Metrics

To evaluate the performance of the proposed method, the Average Percentage of Bad Pixels (APBP) is defined as follows:

$$APBP = \frac{1}{N} \sum_{p \in P} \mathbb{I} \left(|D_{pred}(p) - D_{gt}(p) > T| \right) \times 100 \tag{9}$$

where $D_{pred}(p)$ and $D_{gt}(p)$ are the estimated disparity and the ground-truth disparity at pixel $p \in P$, respectively. T is the error tolerance threshold. P is the total number of valid pixel in the disparity map.

3.3. Performance Evaluation on DrivingStereo Dataset

The DrivingStereo dataset was used to verify the effectiveness of the proposed SACA method in comparison to the SGM. Fig. 4 shows the APBP of the proposed SACA and SGM under various difficult driving conditions on a pair of stereo images. For the case of cloudy conditions, the proposed SACA and SGM achieve an APBP of 3.6% and 12.77%, respectively. For the case of foggy conditions, the proposed SACA and SGM achieve an APBP of 5.01% and 46.65%, respectively. For the case of sunny conditions, the proposed SACA and SGM achieve an APBP of 11.16% and 25.76%, respectively. The results show that the proposed SACA achieved much better performance than SGM due to the benefits of local cross-attention features. Figs. 5, 6, and 7 show the performance of the proposed SACA and SGM on a total of 500 images from the cloudy conditions, 500 images from the foggy conditions, and 447 images from the sunny conditions. For the cloudy case, the proposed SACA reduced the average percentage of bad pixels of the SGM by 2.93 times (from 14.28% to 4.87%). For the foggy case, the proposed SACA reduced the average percentage of bad pixels of the SGM by 2.7 times (from 20.55% to 7.5%). For the sunny case, the proposed SACA reduced the average percentage of bad pixels of the SGM by 2.8 times (from 13.62% to 4.97%). Thus, comprehensive experiments show that our proposed SACA method yields stable and better performance than SGM under various testing conditions using the DrivingStereo dataset.

3.4. Performance Evaluation on HCI Stereo Dataset

To evaluate the performance of the proposed system under more challenging driving conditions, we assess the performance of the proposed method and the original SGM under 4 conditions (Fig. 8): reflecting car, sun-flare, rain flare, and night and snow. The visualization of disparity results confirms that the proposed SACA method still produces better results than the SGM under these challenging test cases. In particular, the performance of the proposed SACA significantly outperforms the performance of SGM under the night and snow conditions.

3.5. Performance Evaluation on KITTI Stereo Dataset

To further evaluate the performance of the proposed system, we conducted more experiments on 194 stereo images from the KITTI training 2012 (Fig. 9) and 200 stereo images from the KITTI training 2015 (Fig.

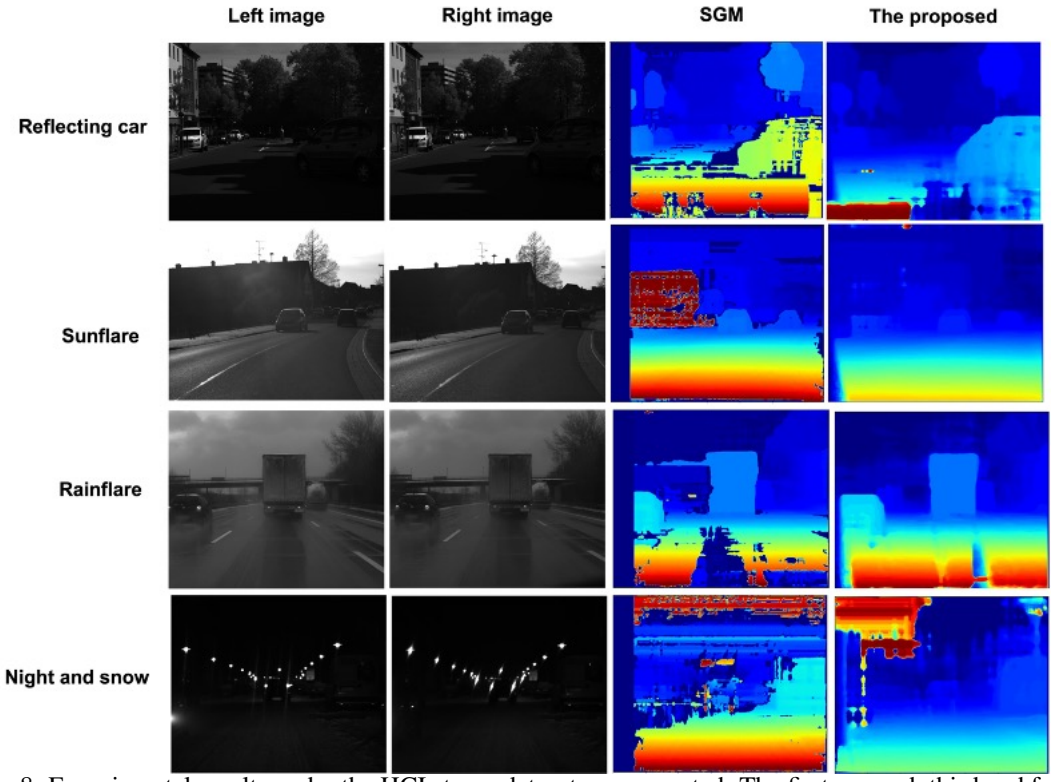


Figure 8. Experimental results under the HCI stereo dataset are presented. The first, second, third and fourth rows show the results of the SGM and the proposed method under reflecting car, sunflare, rain flare and night and snow conditions, respectively.

10) stereo datasets. The proposed SACA method and the SGM achieved average percentages of bad pixels of 4.98% and 12.37%, respectively, under the KITTI 2012 dataset. The proposed SACA method and the SGM achieved average percentages of bad pixels of 5.78% and 12.05%, respectively, under the KITTI 2015 dataset. To verify the performance of the proposed method against other state-of-the-art methods, we conducted more experiments on the KITTI 2012 testing dataset, as shown in Fig. 11. The APBP of the proposed method, SGM, and permutation are 5.50%, 7.68%, and 8.04%, respectively. The experimental results prove that the proposed SACA still achieves comparable performance in comparison to the state-of-the-art non-deep learning and deep learning methods.

To continue evaluating the performance of the proposed method under the most challenging conditions in the KITTI 2012 dataset, we conducted experiments on the reflective regions of 194 testing stereo images. The proposed method, which achieves the lowest Average Pixel-wise Boundary Position (APBP) error of 12.01, outperforming all other compared methods: SGM (12.48), CRD-Fusion (12.87), and Permutation (18.78). The superior accuracy of the proposed method is primarily due to its integration of an attention mechanism for feature extraction, which allows the model to selectively focus on the most informative features, improving spatial detail and boundary delineation. Additionally, the use of evolutionary algorithms for parameter estimation ensures a more adaptive and optimized configuration, enhancing performance across varied input conditions. In contrast, methods like SGM and CRD-Fusion lack this adaptive capability, which limits their precision. Overall, the results highlight the advantage of combining advanced attention-based architectures with intelligent parameter optimization.

3.6. Performance Evaluation to Other Deep Learning Models

To continue evaluating the performance of the proposed method with existing deep-learning based method, we conduce the experiment on the Driving Stereo dataset, under the Foggy conditions. Fig. 12 shows the result of various method. The results presented in the experiment clearly highlight the superiority of the proposed method over all competing approaches across multiple benchmark images. The quantitative evaluation

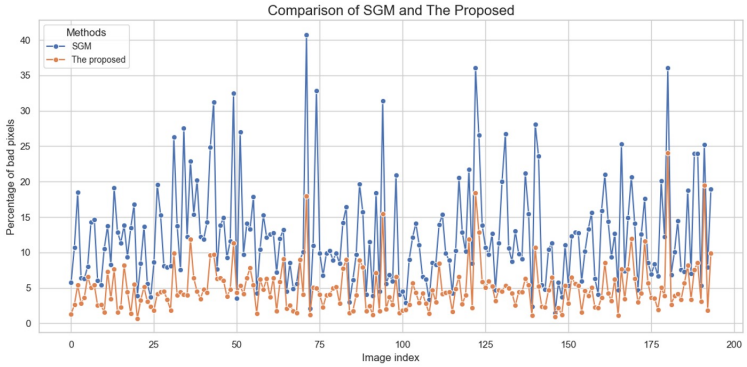


Figure 9. The experimental results of the SGM-based method and the proposed method with 194 training stereo images using the KITTI 2012 dataset.

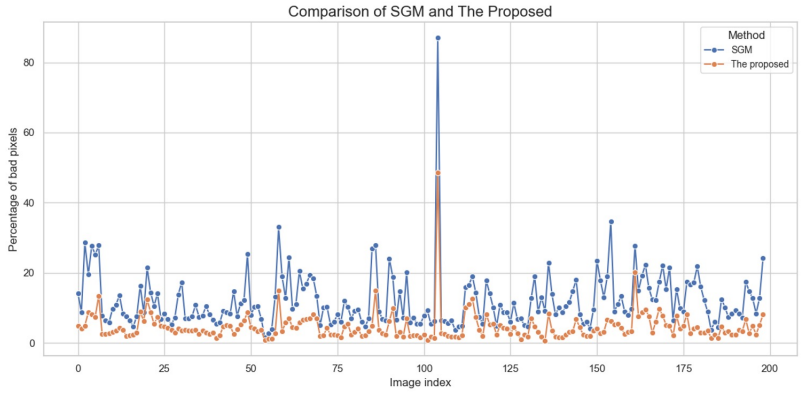


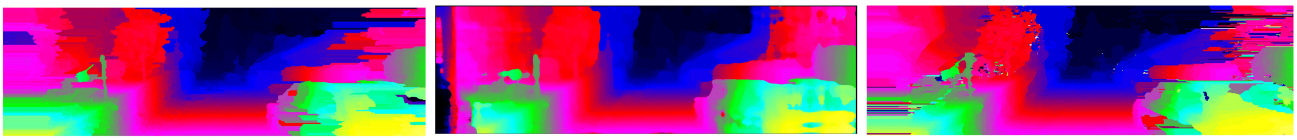
Figure 10. The experimental results of the SGM-based method and the proposed method with 200 training stereo images using the KITTI 2015 dataset.

based on APBP clearly demonstrates the effectiveness of the proposed method compared with both classical stereo approaches and state-of-the-art learning-based techniques. the proposed method achieves the lowest APBP value of 4.65, outperforming traditional SGM-based methods such as SGM (20.86), permutation-based SGM (10.60), m-SGM (13.44), and s-SGM (23.55) by a substantial margin. This confirms that conventional handcrafted cost functions remain highly sensitive to noise and low-texture regions. Additionally, although advanced learning-based approaches such as CRD-Fusion (6.31), FEDL (7.93), and RAPIDFlow (6.97) show improved accuracy, they still perform worse than the proposed method.

The superiority of the proposed method is attributed to two key innovations: (1) the self- and cross-attention mechanism that enhances robustness by capturing reliable contextual relationships between stereo image pairs, and (2) the adaptive optimization of SGM penalties through Genetic Algorithms, ensuring stable performance under diverse environmental variations.

4. RUNNING TIME AND LIMITATIONS

To verify the processing time of the proposed method, the experiment was performed on a computer equipped with a Titan X 12 GB GPU, 8 Intel Xeon 3.5 GHz CPUs, and 8 GB RAM. Our proposed SACA method achieved 0.02s/frame. The proposed SACA does not achieve good results under extremely difficult conditions in the DrivingStereo dataset. The proposed method fails to generate disparity results under heavy rain and low light in tunnels, with the reflecting effects from sun flare.



(a) (b) (c)

Figure 11. (a), (b) and (c) are the APBP results of SGM (7.68%), the proposed method (5.50%), and permutation (8.04%), respectively, on the KITTI 2012 testing dataset.

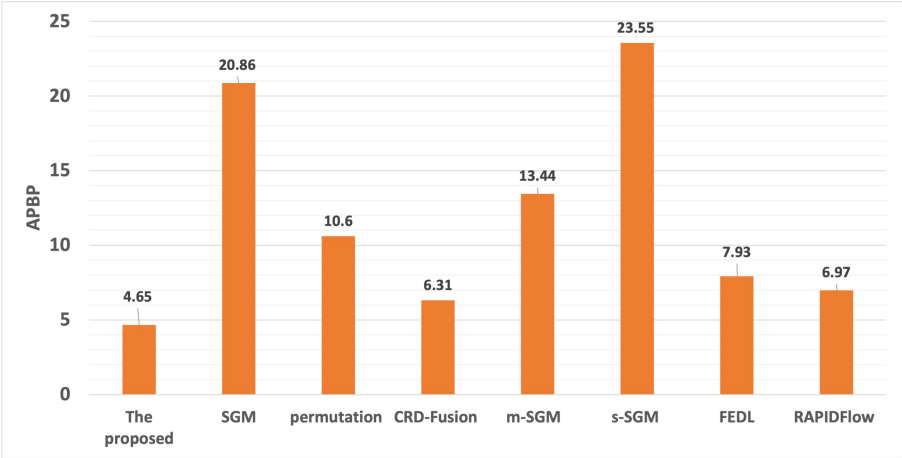


Figure 12. Experimental results of the proposed method and other non-deep learning and deep learning model under Foggy conditions on the Driving Stereo dataset.

5. CONCLUSION

This research introduces an efficient method for a real-time and robust stereo matching system under various and complex driving conditions. The proposed method was designed based on local and cross-attention features, along with an evolutionary algorithm for parameter estimation. Experimental results show that the proposed method achieves stable and robust performance under various testing datasets, including KITTI, HCI, and DrivingStereo. In the future, we aim to investigate the benefits of both deep learning and non-deep learning models to address the current limitations of the proposed method.

ACKNOWLEDGMENTS

We would like to thanks to FPT University, Can Tho Campus, Vietnam for funding this research.

REFERENCES

[1] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1738–1764, 2022. [Online]. Available: https://doi.org/10.1109/TPAMI.2020.3032602

[2] F. Tosi, L. Bartolomei, and M. Poggi, "A survey on deep stereo matching in the twenties," *International Journal of Computer Vision*, 2025. [Online]. Available: https://doi.org/10.1007/s11263-024-02331-0

[3] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 242–16 251. [Online]. Available: https://doi.org/10.1109/CVPR52688.2022.01578

[4] W. Zhong, Y. Yuan, X. Ye, D. Zheng, and R. Xu, "Underwater stereo matching via unsupervised appearance and feature adaptation networks," pp. 2295–2299, 2022. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9746701

[5] H. Dai, X. Zhang, Y. Zhao, H. Sun, and N. Zheng, "Adaptive disparity candidates prediction network for efficient real-time stereo matching," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 3099–3110, 2022. [Online]. Available: https://doi.org/10.1109/TCSVT.2021.3102109

[6] X. Lai, Y. Zhang, and Z. Li, "An improved stereo matching algorithm based on joint similarity measure and adaptive weights," *Applied Sciences*, vol. 13, no. 1, p. 514, 2023. [Online]. Available: https://doi.org/10.3390/app13010514

[7] G. Xu, X. Wang, X. Ding, and X. Yang, "Iterative geometry encoding volume for stereo matching," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 21 919–21 928. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.02099

[8] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2024, pp. 19 701–19 710. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.01863

[9] K. Li, L. Wang, Y. Zhang, K. Xue, S. Zhou, and Y. Guo, "Los: Local structure-guided stereo matching," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 19 746–19 756. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.01867

[10] Z. Cheng, J. Yang, and H. Li, "Stereo matching in time: 100+ fps video stereo matching for extended reality," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, HI, USA: IEEE, 2024, pp. 8704–8713. [Online]. Available: https://doi.org/10.1109/WACV57701.2024.00852

[11] X. Yue *et al.*, "Semi-stereo: A universal stereo matching framework for imperfect data via semi-supervised learning," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, 2024, pp. 646–655. [Online]. Available: https://doi.org/10.1109/CVPRW63382.2024.00069

[12] O.-H. Kwon and E. Zell, "Image-coupled volume propagation for stereo matching," in *2023 IEEE International Conference on Image Processing (ICIP)*. Kuala Lumpur, Malaysia: IEEE, 2023, pp. 2510–2514. [Online]. Available: https://doi.org/10.1109/ICIP49359.2023.10222247

[13] J. Song, Q. Zhu, J. Lin, and M. Ghaffari, "BDIS: Bayesian Dense Inverse Searching Method for Real-Time Stereo Surgical Image Matching," *IEEE Transactions on Robotics*, vol. 39, no. 2, pp. 1388–1406, Apr. 2023. [Online]. Available: https://doi.org/10.1109/TRO.2022.3215018

[14] Z. Quan, B. Wu, and L. Luo, "An image stereo matching algorithm with multi-spectral attention mechanism," *Sensors*, vol. 23, no. 19, p. 8179, 2023. [Online]. Available: https://doi.org/10.3390/s23198179

[15] Y. Zhang, M. Poggi, and S. Mattoccia, "Temporalstereo: Efficient spatial-temporal stereo matching network," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Detroit, MI, USA: IEEE, 2023, pp. 9528–9535. [Online]. Available: https://doi.org/10.1109/IROS55552.2023.10341598

[16] W. Gao, Y. Cai, Y. Akoudad, Y. Yang, and J. Chen, "Ma-stereo: Real-time stereo matching via multi-scale attention fusion and spatial error-aware refinement," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9954–9961, Nov. 2024. [Online]. Available: https://doi.org/10.1109/LRA.2024.3468173

[17] X. Ye, J. Li, H. Wang, H. Huang, and X. Zhang, "Efficient stereo matching leveraging deep local and context information," *IEEE Access*, vol. 5, pp. 18 745–18 755, 2017. [Online]. Available: https://doi.org/10.1109/ACCESS.2017.2754318

[18] B. Zhang, X. Wu, S. Liu, and W. Wan, "A fast stereo matching network based on multi-scale feature fusion," in *2024 IEEE 7th International Conference on Electronic Information and Communication Technology (ICEICT)*. Xi'an, China: IEEE, 2024, pp. 491–496. [Online]. Available: https://doi.org/10.1109/ICEICT61637.2024.10670969

[19] X. Yang, L. He, Y. Zhao, H. Sang, Z. L. Yang, and X. J. Cheng, "Multi-attention network for stereo matching," *IEEE Access*, vol. 8, pp. 113 371–113 382, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3003375

[20] X. Wang, J. Yu, Z. Sun, J. Sun, and Y. Su, "Multi-scale graph neural network for global stereo matching," *Signal Processing: Image Communication*, vol. 118, p. 117026, 2023. [Online]. Available: https://doi.org/10.1016/j.image.2023.117026

[21] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, pp. 1956–1965. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00203

[22] X. Su, S. Liu, R. Li, Z. Bing, and A. Knoll, "Efficient stereo matching using swin transformer and multilevel feature consistency in autonomous mobile systems," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 5, pp. 7957–7965, May 2024. [Online]. Available: https://doi.org/10.1109/TII.2024.3367033

[23] X. Yang, Y. Zhao, Z. Feng, H. Sang, Z. Zhang, G. Zhang, and L. He, "A light-weight stereo matching network based on multi-scale features fusion and robust disparity refinement," *IET Image Processing*, vol. 17, pp. 1797–1811, 2023. [Online]. Available: https://doi.org/10.1049/ipr2.12756

[24] J. Qi and L. Liu, "The stereo matching algorithm based on an improved adaptive support window," *IET Image Processing*, vol. 16, no. 10, pp. 2803–2816, Oct 2022. [Online]. Available: https://doi.org/10.1049/ipr2.1234567

[25] Y. Zhang, X. Li, and Z. Li, "Lidar-guided stereo matching with a spatial consistency constraint," *arXiv preprint arXiv:2202.09953*, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2202.09953

[26] S. Yang, H. Chen, and W. Chen, "Generalized stereo matching method based on iterative optimization of hierarchical graph structure consistency cost," *Remote Sensing*, vol. 15, no. 9, p. 2369, 2023. [Online]. Available: https://doi.org/10.3390/rs15092369

[27] W. Iqbal, J. A. Paffenholz, and M. Mehltretter, "Guiding deep learning with expert knowledge for dense stereo matching," *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, vol. 91, pp. 365–380, 2023. [Online]. Available: https://doi.org/10.1007/s41064-023-00252-0

[28] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1738–1764, Apr. 2022. [Online]. Available: https://doi.org/10.1109/TPAMI.2020.3032602

[29] G. Suh, J. Lim, S. Jeong, and T. Park, "A real-time design of stereo vision system using disparity calibration," in *2024 Fifteenth International Conference on Ubiquitous and Future Networks (ICUFN)*. Budapest, Hungary: IEEE, 2024, pp. 532–537. [Online]. Available: https://doi.org/10.1109/ICUFN61752.2024.10625225

[30] H. Wenbin and S. Kamata, "Unsupervised learning for stereo depth estimation using efficient correspondence matching," in *Proceedings of the 5th International Conference on Advanced Image Processing (ICAIP '21)*. New York, NY, USA: ACM, 2022, pp. 30–34. [Online]. Available: https://doi.org/10.1145/3502827.3502828

[31] K. Wegner and O. Stankiewicz, "Depth estimation using modified cost function for occlusion handling," *Signal, Image and Video Processing*, vol. 13, pp. 1539–1547, 2019. [Online]. Available: https://doi.org/10.1007/s11760-019-01487-4

[32] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, "Object-centric stereo matching for 3d object detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE, 2020, pp. 8383–8389. [Online]. Available: https://doi.org/10.1007/s11760-019-01487-4

[33] S. Jeon and Y. S. Heo, "Efficient multi-scale stereo-matching network using adaptive cost volume filtering," *Sensors*, vol. 22, no. 15, p. 5500, 2022. [Online]. Available: https://doi.org/10.3390/s22155500

[34] E.-T. Baek and H. J. Yang, "Cost aggregation for stereo matching using total generalized variation with fusion tensor," *IEEE Access*, vol. 7, pp. 134 505–134 513, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2941215

[35] H. Liu *et al.*, "Stereo matching algorithm based on two-phase adaptive optimization of ad-census and gradient fusion," in *2021 IEEE International Conference on Real-time Computing and Robotics (RCAR)*. Xining, China: IEEE, 2021, pp. 726–731. [Online]. Available: https://doi.org/10.1109/RCAR52367.2021.9517511

[36] W. Chen, H. Chen, and S. Yang, "Self-supervised stereo matching method based on srwp and pcam for urban satellite images," *Remote Sensing*, vol. 14, no. 7, p. 1636, 2022. [Online]. Available: https://doi.org/10.3390/rs14071636

[37] B. Liu, K. Chen, S.-L. Peng, and M. Zhao, "Adaptive aggregate stereo matching network with depth map super-resolution," *Sensors*, vol. 22, no. 12, p. 4548, 2022. [Online]. Available: https://doi.org/10.3390/s22124548

[38] S. Meister, B. Jähne, and D. Kondermann, "A stereo camera system for the generation of real-world benchmark datasets with ground truth," 2011. [Online]. Available: https://doi.org/10.11588/heidok.00011610

[39] M. Jang, H. Yoon, S. Lee, J. Kang, and S. Lee, "A comparison and evaluation of stereo matching on active stereo images," *Sensors*, vol. 22, no. 9, p. 3332, May 2022. [Online]. Available: https://doi.org/10.3390/s22093332

[40] M. J. A. Dujaili, "Survey on facial expressions recognition: Databases, features, and classification schemes," *Multimedia Tools and Applications*, vol. 83, pp. 7457–7478, 2024. [Online]. Available: https://doi.org/10.1007/s11042-023-15139-w

[41] V. D. Nguyen, H. V. Nguyen, and J. W. Jeon, "Robust stereo data cost with a learning strategy," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 2, pp. 248–258, 2017. [Online]. Available: https://doi.org/10.1109/TITS.2016.2563661

[42] V. D. Nguyen, T. T. Nguyen, D. D. Nguyen, S. J. Lee, and J. W. Jeon, "A fast evolutionary algorithm for real-time vehicle detection," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 6, pp. 2453–2468, Jul 2013. [Online]. Available: https://doi.org/10.1109/TVT.2013.224291

[43] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 899–908. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00099

[44] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361. [Online]. Available: https://doi.org/10.1109/CVPR.2012.6248074

[45] C. Deng, D. Liu, H. Zhang, J. Li, and B. Shi, "Semi-global stereo matching algorithm based on multi-scale information fusion," *Applied Sciences*, vol. 13, no. 2, p. 1027, Jan 2023. [Online]. Available: https://doi.org/10.3390/app13021027

[46] I. Firmansyah, B. Setiadi, R. Indrawijaya, A. S. Satyawan, D. Kurniawan, and A. A. Waskita, "An efficient hardware implementation of fpga-based stereo matching using a single path sgm direction," pp. 400–404, 2024. [Online]. Available: https://doi.org/10.1109/IC3INA64086.2024.10732374

[47] V. Q. Dinh, P. H. Nguyen, and V. D. Nguyen, "Feature engineering and deep learning for stereo matching under adverse driving conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 7855–7865, Jul 2022. [Online]. Available: https://doi.org/10.1109/TITS.2021.3073557

[48] P.-A. Rousseau and S. Roy, "A permutation model for the self-supervised stereo matching problem," pp. 122–131, 2022. [Online]. Available: https://doi.org/10.1109/CRV55824.2022.0002

[49] X. Fan, S. Jeon, and B. Fidan, "Occlusion-aware self-supervised stereo matching with confidence guided raw disparity fusion," in *2022 19th Conference on Robots and Vision (CRV)*, 2022, pp. 132–139. [Online]. Available: https://doi.org/10.1109/CRV55824.2022.00025

[50] H. Morimitsu, X. Zhu, R. M. Cesar, X. Ji, and X.-C. Yin, "Rapidflow: Recurrent adaptable pyramids with iterative decoding for efficient optical flow estimation," pp. 2946–2952, 2024. [Online]. Available: https://doi.org/10.1109/ICRA57147.2024.10610277

BIOGRAPHIES OF AUTHORS

Vinh Dinh Nguyen received the B.Sc. (magna cum laude) degree in computer science from Nong Lam University, Ho Chi Minh City, Vietnam, in 2007, and the M.S. and Ph.D. degrees in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2012 and 2015, respectively.
From 2015 to 2019 he was a research professor at College of Information and Computer Engineering, Sungkyunkwan University, South Korea. From 2019 to 2021 he was a researcher and lecturer at School of Information and Technology, Eastern Internatinal Uninversity, Vietnam. Since 2021, he has been with the Department of Information Technology, FPT University, Can Tho, Vietnam. His research interests include deep learning, computer vision, and machine learning. He can be contacted at email: vinhnd29@fe.edu.vn

Nhan Huu Tran is currently a bachelor's student at FPT University, Can Tho, Vietnam. His research interests include deep learning, computer vision, and machine learning. He can be contacted at email: NhanTHCE181655@fpt.edu.vn