2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Hybrid Stereo Dense Depth Estimation for Robotic Tasks in Industrial Automation

Suhani Singh, Michael Suppa, Raúl Suárez and Jan Rosell

***Abstract*— We introduce a simple yet effective approach for dense depth reconstruction that operates directly on raw disparity data, eliminating the need for additional disparity refinement stages. By leveraging disparity maps generated from conventional stereo methods, we train a U-Net-based model to directly map disparity to depth, bypassing complex feature engineering. Our method capitalizes on the robustness of traditional stereo matching techniques to varying scenes, focusing exclusively on dense depth reconstruction. This approach not only simplifies the training process but also significantly reduces the requirement for large-scale training datasets. Extensive evaluations demonstrate that our method surpasses classical stereo matching frameworks and state-of-the-art classical post-refinement techniques, achieving superior accuracy. Additionally, our approach offers competitive inference times, comparable to classical as well as end-to-end deep learning methods, making it highly suitable for real-time robotic applications.**

I. INTRODUCTION

Depth perception is a fundamental aspect of robotic systems operating in dynamic and unstructured environments, particularly in industrial settings where precise object manipulation is crucial for tasks such as pick-and-place operations. Stereo vision, leveraging the disparities between images captured by a stereo camera, presents a promising avenue for depth estimation in such scenarios. By translating pixel disparities into accurate depth maps, stereo vision facilitates enhanced spatial understanding critical for robotic decision-making and manipulation tasks.

Efficiently generating dense correspondences between stereo image pairs for depth perception poses a critical challenge in computer vision— particularly in resource-constrained environments. Existing methods, whether traditional algorithms or deep learning-based approaches, grapple with the delicate balance between precision and computational efficiency. The widespread adoption of stereo vision in industrial automation and mobile devices, such as autonomous cars [1] and unmanned aerial vehicles [2], has been propelled by recent advancements in fully-featured embedded microcomputers [3]. However, the evolving landscape underscores the pressing need for stereo matching solutions capable of navigating this balance adeptly.

Suhani Singh and Michael Suppa are with Roboception GmbH, Munich 81241, Germany. {suhani.singh, michael.suppa}@roboception.de

Raúl Suárez and Jan Rosell are with the Institut d'Organització i Control de Sistemes Industrials (IOC), Universitat Politècnica de Catalunya (UPC), 08028 Barcelona, Spain. {raul.suarez, jan.rosell}@upc.edu

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie-Sklodowska Curie grant agreement No. 956670 and is part of the 5GSMARTFACT MSCA-ITN project.

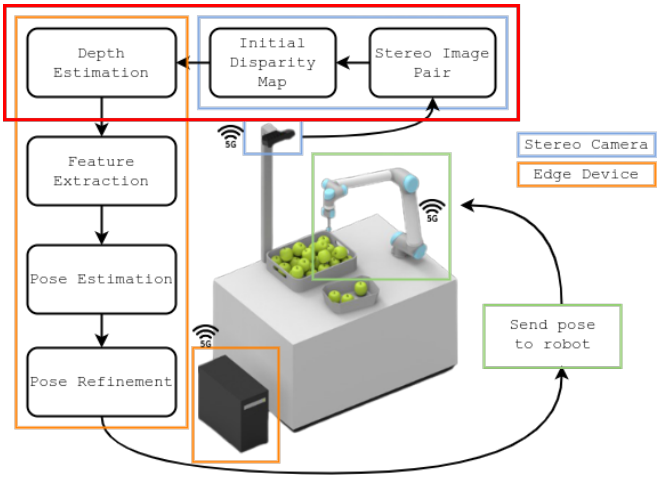

Fig. 1. Depth Estimation on Edge: The proposed framework (highlighted in red) embedded in a robotic picking task. Our approach uses the raw disparity image of a classical stereo method as input for generating a dense depth map for object localisation

To generate dense depths from initial disparity maps, refinement techniques are commonly employed in post-processing stages. Traditional approaches, such as those utilizing left-to-right consistency checks (LRC) [4], often incorporate local refinement strategies like median filters [5], [6]. However, these methods introduce significant computational overhead, leading to increased processing times. In contrast, deep-learning based methods, renowned for their high accuracy in disparity estimation and refinement, are hampered by slow inference times. The application of time-consuming 3D convolutions on a 4D feature volume further escalates the computational cost [7]. Moreover, while direct regression methods for disparity estimation exhibit effectiveness in scenarios with abundant training data and similar train-test distributions, their generalization capability remains limited.

Within the scope of industrial automation, the effectiveness of deep learning-based approaches hinges on the availability of high-quality training datasets containing stereo images of 3D objects paired with accurate ground truth depth maps. However, there remains a notable scarcity of datasets [8] tailored to the specific requirements of stereo-based depth estimation for objects industrial environments, posing a significant challenge for researchers and practitioners alike.

To overcome these challenges, our proposed approach provides a dual benefit. Firstly, it significantly reduces computational complexity by eliminating the need for dedicated

disparity refinement, thereby streamlining the direct depth estimation process from disparity data. This optimization reduces the training data requirements by half, as the model no longer needs both left and right images during training—only the disparity data is required. Secondly, by harnessing edge computing capabilities, our approach offers an effective solution for real-world applications with limited computational resources, as demonstrated in Fig. 1. Transmitting locally generated disparity data to the edge via wireless networks (e.g., 5G) is significantly more efficient than transmitting stereo image pairs due to the reduced data size. Moreover, the utilization of 5G ensures minimal transmission latency, further enhancing processing speed, which is critical for time-sensitive applications. This combination of reduced data volume and low-latency processing makes our solution particularly well-suited for dynamic environments where real-time decision-making is essential.

Furthermore, our architecture integrates smoothly with existing disparity estimation techniques, acting as a lightweight augmentation for producing dense depth reconstructions. By relying solely on disparity data for training, we achieve a significant reduction in both data and computational resource demands, further optimizing the efficiency of the system without the need for supplementary inputs.

II. RELATED WORKS

A. Depth through Conventional Stereo

Depth reconstruction through traditional stereo algorithms takes $n = 2$ rectified images as input and aims to compute the disparity of each pixel by matching the pixels along corresponding epipolar lines, which enables depth estimation via triangulation. [9] formulates this problem as minimizing an energy function $E(D)$, formulated as:

$$E(D) = \sum_{x} C(x, dx) + \sum_{x} \sum_{y \in N_x} E_s(dx, dy)$$

Here, x and y represent image pixels, where N_x is the set of pixels within the neighborhood of x. The first term of the equation represents the matching cost whereas the second term is a regularization criteria for constraints like smoothness and left-right consistency. In rectified stereo pairs, $dx = D(x) \in [d_{\text{min}}, d_{\text{max}}]$ determines depth through triangulation. Discretizing disparity into n_d levels creates a 3D cost volume of size $W \times H \times n_d$. For multi view stereo ($n \geq 2$), $C(x, dx)$ gauges the inverse likelihood of x having depth dx on the reference image.

A traditional stereo algorithm typically follows a combination of the following steps [10], [11] - (1) Matching cost computation,(2) Cost aggregation, (3) Disparity computation, and (4) Disparity refinement. Following this pipeline, algorithms are categorized as local, global, or semi-global. Local algorithms ascertain disparity by identifying the lowest cost or highest correlation through a winner-takes-all (WTA) strategy. Global stereo matching algorithms conceptualize the disparity estimation problem as a global energy minimization challenge, often addressed using optimization algorithms based on Markov random fields (MRF) [12], such as graph-cut (GC) [13] and dynamic programming (DP) [14]. Semi-global matching (SGM) [15] approximates MRF inference by aggregating costs in all directions within the image. This approach significantly enhances the balance between the accuracy and efficiency of stereo matching.

B. End-to-end Depth Estimation

Deep learning based works address the stereo matching challenge through an end-to-end trained pipeline. Initial methods in this domain [16], [17] employ a single encoder-decoder architecture. This architecture combines the left and right images into a 6D volume and predict the disparity map. Despite their runtime efficiency, these methods demand a substantial volume of training data. Recent approaches [18], [19] replicate the conventional stereo matching pipeline, fragmenting the training process into differentiable blocks, thereby facilitating end-to-end training. In spite of achieving impressive results, these methods have a huge number of learnable parameters resulting in long inference time [3]. Supervised approaches in this domain have demonstrated remarkable success in disparity estimation. However, the significant amount of ground truth data required for training poses a time-consuming and labor-intensive challenge. To circumvent this, unsupervised stereo methods [20], [21] have emerged to eliminate the dependency on disparity ground truth. While supervised and unsupervised approaches share architectural similarities, their training processes differ significantly. Unsupervised methods, despite recent advancements, still struggle with performance instability in challenging areas, especially those with occlusions, due to the sensitivity of a single network to outliers. Consequently, a noticeable performance gap persists between existing supervised and unsupervised approaches.

C. Disparity Refinement

Disparity maps often exhibit inherent noise and inaccuracies needing meticulous refinement. In [22], a novel neural architecture for disparity refinement was introduced with a specific focus on advancing 3D computer vision capabilities on consumer-grade devices. Within the domain of medical imaging, [23] addressed the intricacies of laparoscopic images by devising a sophisticated disparity refinement framework tailored for learning-based stereo matching methods. This framework incorporates both local and global disparity refinement strategies, demonstrating efficacy in refining noise-corrupted disparity maps without compromising prediction accuracy. In tackling the challenge of reliable matching in weakly matchable regions, [24] proposed a stereo matching network that adopts a pixel-wise matchability perspective. This network performs regression on both disparity and matchability maps, leveraging 3D probability volumes. Additionally, a matchability-aware disparity refinement module is introduced to augment depth inference.

Lastly, [25] introduces a recurrent network tailored for disparity refinement by integrating recurrence, multi-scale processing, and residual design to enhance input disparity

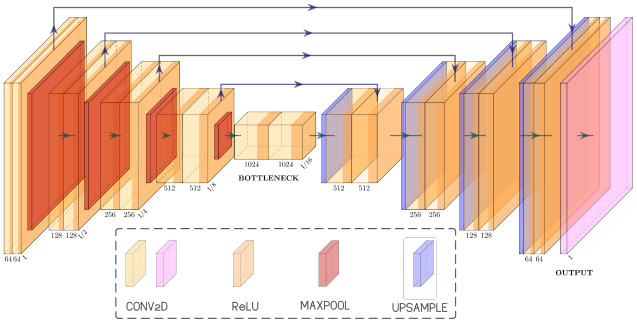

Fig. 2. Network Architecture: Leveraging pre-generated disparity data, the proposed network based on U-Net [26] adeptly learns disparity to depth mapping for a comprehensive depth estimation.

maps. The method demonstrates substantial error rate reductions, particularly on Multi-Channel CNN (MC-CNN)-generated [5] disparity maps. While promising, the approach has limitations, as it lags slightly behind complete end-to-end stereo pipelines in terms of accuracy.

III. METHOD

Our objective is to develop a robust and computationally efficient method for mapping raw disparity to dense depth. Instead of relying on an end-to-end network architecture that simultaneously processes stereo image pairs to produce disparity or depth information, our approach capitalizes on pre-computed disparity maps generated directly by a stereo sensor. The stereo sensor captures left and right images, computes the initial disparity map on-board, and transmits this map to the edge device for further refinement. This approach not only streamlines the computational process but also ensures that our methodology can be seamlessly integrated into existing stereo vision systems, regardless of whether they are based on traditional or deep learning principles.

A. Network Architecture

We leverage a U-Net [26] based architecture tailored for depth estimation task, see Fig: 2. The network follows a symmetrical encoder-decoder structure, incorporating skip connections to facilitate feature fusion. The encoder is composed of four consecutive encoder blocks and a bottleneck block, each comprising two convolutional layers with batch normalization and ReLU activation, followed by max-pooling for spatial dimension reduction. Conversely, the decoder consists of four decoder blocks, featuring transposed convolutional layers and two additional convolutional layers with corresponding batch normalization and ReLU activation. The final output layer employs just a convolutional layer, producing a depth map.

B. Disparity as Input

Our network is designed to operate directly on pre-computed disparity maps, which encapsulate critical depth and local structural information. This approach eliminates the need for full stereo image pairs during training, drastically reducing the data and computational complexity compared to end-to-end stereo pipelines. By concentrating on disparity data, the learning process becomes more efficient, with the model focusing on refining depth representations instead of handling raw image inputs. The result is a streamlined training process that requires fewer resources while still achieving high-quality depth predictions, even under challenging conditions such as occlusions, varying lighting, and complex surface textures.

C. Disparity Refinement Process

The core of our methodology lies in refining the disparity maps received from the stereo sensor. The raw disparity maps, which may suffer from noise and inconsistencies due to texture-less regions or large depth discontinuities, are fed into our network for refinement. The U-Net architecture, with its skip connections and deep convolutional layers, is well-suited for this task, enabling the network to effectively enhance the initial disparity maps by improving object boundary delineation, reducing artifacts, and producing smoother depth transitions. The final output is a dense depth map that is more accurate and reliable than the raw disparity input, making it suitable for applications such as robotic manipulation where precision is crucial.

IV. EXPERIMENTATION

The scope of the proposed work is object-detection in industrial and lab automation. Current applications in this domain typically rely on disparities generated by traditional stereo algorithms, subsequently refined to yield dense depth maps. Alternatively, neural networks trained on CAD (Computer-Aided Design) models of target objects are employed for detection purposes. Our concept serves as a simple yet effective augmentation to existing disparity estimation pipelines. By directly generating dense depth maps, the reconstruction achieves sufficient density, facilitating the accurate creation of point clouds for subsequent robotic manipulation tasks.

A. Dataset Preprocessing

The disparity data utilized during training is derived from the left-right image pairs sourced from Falling Things (FAT) [8] dataset, chosen for its relevance to our application. This dataset offers diverse scenes featuring 3D objects set against various backgrounds and lighting conditions, providing a realistic representation of industrial settings. Notably, the FAT dataset is unique in providing valid stereo pairs for all objects in the scene, making it an ideal choice for our proposed network.

The intermediate disparity maps that serve as an input for training are generated using the Semi-Global Matching (SGM) algorithm [15]. Our specific implementation of SGM excludes any disparity refinement modules, ensuring no additional refinement of the SGM-derived disparities. The choice of SGM over alternative algorithms [12], [13], [14], stems from its effectiveness in producing globally consistent disparity maps and our preference for using a classical

algorithm as a base. Despite its robust performance, it is crucial to acknowledge the limitations of SGM, particularly in scenarios with texture-less regions or large depth discontinuities. Irrespective of the choice of our base algorithm, the proposed network should work with disparities generated with any algorithm of choice. Given that the generated disparity is left-aligned, the associated ground truth depth map from the dataset is utilized for supervision during the training process.

B. Model Training

The disparity-depth pairs, serving as the input and label, are first resized to dimensions of *(256, 256)* and then combined into a multi-channel input format. This consolidated input is subsequently fed into the network for processing. For optimizing depth estimation, we utilize the L1 loss, which quantifies the absolute discrepancies between predicted and ground truth depth values. The optimization strategy relies on the Adam optimizer [27], with specific parameter settings, notably betas set to 0.9 and 0.999, and epsilon set to 1e-08. We trained the model for 200 epochs on an NVIDIA GeForce RTX 4090 GPU with a batch size of 16, employing a learning rate scheduler that reduced the learning rate by a factor of 0.1 every 25 epochs. Dropout was applied after the encoder layers during training to prevent overfitting.

V. RESULTS AND DISCUSSION

In this section, we provide an in-depth analysis of the performance of our proposed dense depth estimation framework. We assess our method's accuracy, efficiency, and robustness across various scenarios and compare it with state-of-the-art methods. Both quantitative and qualitative results are discussed, along with the implications for real-world applications in industrial automation.

A. Quantitative Evaluation

When evaluating the enhanced depth, we employ four widely used metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR). These metrics facilitate per-pixel evaluation by comparing the refined depth generated by our method against the ground truth depth.

Mean Absolute Error (MAE) quantifies the average magnitude of errors without considering their direction, offering a straightforward measure of the model's accuracy.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root Mean Square Error (RMSE) provides a more nuanced understanding of performance by additionally penalizing larger errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

TABLE I
COMPREHENSIVE QUANTITATIVE ASSESSMENT OF DEPTH ESTIMATION METHODS

Method	MAE	RMSE	SSIM	PSNR (dB)
Raw SGM Depth	0.849	0.855	0.712	19.5
SGM + Refinement Depth	0.613	0.642	0.765	21.8
Ours	**0.368**	**0.374**	**0.812**	**24.6**

Structural Similarity Index (SSIM) evaluates the similarity between the predicted depth maps and the ground truth by assessing changes in structural information, luminance, and contrast. Unlike MAE and RMSE, SSIM focuses on the preservation of structural integrity, such as edges and textures, which are crucial for accurate depth representation. This metric is particularly effective in capturing perceptual differences that are more aligned with human visual assessment.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μ_x and μ_y are the mean values of the predicted and ground truth depth maps, respectively, σ_x^2 and σ_y^2 are their variances, and σ_{xy} is the covariance between the two. C_1 and C_2 are small constants included to stabilize the division with weak denominators.

Peak Signal-to-Noise Ratio (PSNR) measures the peak error between the predicted depth maps and the ground truth. PSNR is expressed in decibels (dB) and provides an indication of the overall quality of the reconstructed depth map. Higher PSNR values indicate better quality, as they correspond to lower levels of distortion. PSNR is particularly useful for assessing the fidelity of the depth maps in terms of preserving the overall intensity range and minimizing noise.

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

where MAX_I is the maximum possible pixel value of the depth map, and MSE is the Mean Squared Error between the predicted and ground truth depth maps.

In Table I, we assess the losses on depth generated from three methods: (a) Raw SGM Only Depth; (b) SGM with Post Refinement Depth; and (c) Depth from our method which refines the raw SGM disparity.

Our method consistently outperforms the baselines across all metrics, reflecting its superior depth estimation capabilities. The significantly lower MAE of **0.368** and RMSE of **0.374** achieved by our approach indicate enhanced accuracy and consistency in depth predictions, which is crucial for high-precision applications such as robotic manipulation and autonomous navigation. Furthermore, the higher SSIM score of **0.812** demonstrates that our method better preserves structural information, such as edges and contours, in the depth maps, which is vital for tasks like object recognition and scene understanding. Lastly, the increased PSNR value of **24.6 dB** suggests that our method effectively reduces noise

TABLE II

COMPARISON OF INFERENCE TIMES FOR DIFFERENT METHODS

Method	Time (s)
SGM [15]	0.049
MoCha-Stereo [28]	0.27
GANet+ADL [29]	0.67
Selective-IGEV [30]	0.24
MC-Stereo [31]	0.40
Combined Pipeline (SGM + Our Method)	**0.158**

and enhances the overall quality of the depth maps, making them more reliable for downstream applications.

Our proposed approach serves as an augmentation to existing disparity estimation pipelines, making traditional benchmarking against standalone depth estimation methods less directly applicable. Designed for seamless integration with established stereo matching systems, our method enhances their capabilities without requiring a complete overhaul. Conventional metrics used for benchmarking may not fully capture the benefits of our approach, which focuses on enhancing existing systems and addressing industrial challenges. Therefore, while benchmarking against existing methods remains important, we advocate for a nuanced assessment considering our method's role as an augmentation to established pipelines.

Lastly, our experiments focused on evaluating the inference time of our network as shown in Table II. While the SGM stereo method showed inference times of approximately 0.049s, our method demonstrated an average inference time of approximately **0.109s** on an NVIDIA GeForce RTX 4090 GPU. This inference time suggests that the combined hybrid pipeline of SGM disparity with our proposed depth estimation would take around 0.158 seconds, significantly less than most deep learning-based stereo disparity methods [9]. The expedited inference time emphasizes the suitability of our approach for real-time applications, where rapid and efficient depth estimation is crucial.

B. Visual Inspection

To further validate the effectiveness of our approach, we performed a qualitative assessment by visually comparing the depth maps generated by different methods. Fig. 3 showcases the depth maps produced by Raw SGM, SGM + Post-Refinement, and SGM + our proposed method.

One of the key challenges in depth estimation is accurately capturing depth information at object boundaries and in occluded regions. As shown in Fig. 3, our method significantly improves the clarity and sharpness of object boundaries compared to the baseline methods. The post-refinement process in traditional SGM often introduces artifacts and fails to handle occlusions effectively, leading to inconsistencies in the depth map. In contrast, our approach produces smoother transitions and more consistent depth values across occluded regions, which is critical for applications like collision avoidance and grasping in robotics.

Another improvement offered by our method is the reduction of noise in the depth maps. The raw SGM disparity maps often contain noise, particularly in texture-less areas or regions with weak stereo correspondences. Our model effectively filters out these noise artifacts, resulting in cleaner and more continuous depth maps. This enhancement is particularly beneficial in scenarios where accurate 3D modeling is required, such as in industrial inspection or augmented reality.

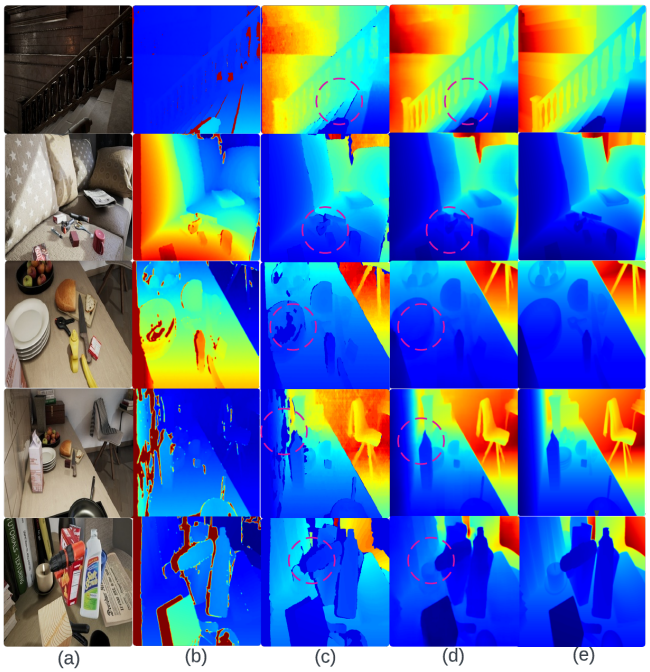

Fig. 3. (a) Left Image; (b) Raw SGM Disparity (Input); (c) Depth from SGM + Disparity Refinement; **(d) SGM + Ours**; (e) Ground Truth Depth; The raw disparity input represents the initial data fed into our model. The depth map generated using SGM with post-processing exhibits noise and inconsistencies typical of the original disparity. In contrast, our method significantly refines object boundaries and depth discontinuities, resulting in a more dense depth reconstruction.

C. Limitations of Dataset Availability

Our depth estimation study faced a notable challenge due to the limited availability of diverse stereo 3D object datasets with paired depth maps [8]. We relied on the FAT dataset for experimentation, as it met our criteria, yet it doesn't encompass the full spectrum of real-world scenes and object configurations. Consequently, our testing scope was confined to this dataset, limiting the generalizability of our findings. Nonetheless, our results offer valuable insights into our model's performance in depth estimation tasks. Moving forward, there's a pressing need for more comprehensive and varied datasets containing stereo pairs with depth maps to enable more thorough evaluations of stereo depth estimation algorithms.

VI. CONCLUSION

In this paper, we introduced a novel approach for dense depth reconstruction directly from raw disparity data, eliminating the need for a dedicated disparity refinement stage.

Leveraging locally generated raw disparities, our U-Net-based model learns disparity-to-depth mapping directly. Our method demonstrated superior performance compared to baseline methods through comprehensive evaluation. Visual comparisons highlighted significant enhancements in object boundary delineation, artifact reduction, and occlusion handling. Our method is optimized for edge deployment alongside any stereo sensor, as it solely requires a disparity map for refinement rather than a stereo pair to generate depth from scratch. This characteristic simplifies implementation and resource utilization, making it a practical solution for real-time depth estimation tasks without significant hardware upgrades. However, limitations in dataset availability emphasize the necessity for more diverse datasets to enable comprehensive evaluations. Future work in this field would involve integrating our model into an existing disparity estimation and object detection pipeline to facilitate 3D object detection in pick-and-place experiments within an industrial setting.

REFERENCES

[1] R. Fan, H. Wang, P. Cai, J. Wu, M. J. Bocus, L. Qiao, and M. Liu, "Learning collision-free space detection from stereo images: Homography matrix brings better data augmentation," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 1, pp. 225–233, 2022.

[2] R. Duan, D. P. Paudel, C. Fu, and P. Lu, "Stereo orientation prior for uav robust and accurate visual odometry," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 5, pp. 3440–3450, 2022.

[3] C.-W. Liu, H. Wang, S. Guo, M. J. Bocus, Q. Chen, and R. Fan, *Stereo Matching: Fundamentals, State-of-the-Art, and Existing Challenges*. Singapore: Springer Nature Singapore, 2023, pp. 63–100.

[4] S. Zhu, Z. Wang, X. Zhang, and Y. Li, "Edge-preserving guided filtering based cost aggregation for stereo matching," *Journal of Visual Communication and Image Representation*, vol. 39, pp. 107–119, 2016.

[5] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2287–2318, jan 2016.

[6] S. Wen, "Convolutional neural network and adaptive guided image filter based stereo matching," in *2017 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2017, pp. 1–6.

[7] Y. Zhong, C. Loop, W. Byeon, S. Birchfield, Y. Dai, K. Zhang, A. Kamenev, T. Breuel, H. Li, and J. Kautz, "Displacement-invariant cost computation for stereo matching," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1196–1209, 2022.

[8] J. Tremblay, T. To, and S. Birchfield, "Falling things: A synthetic dataset for 3D object detection and pose estimation," in *CVPR Workshop on Real World Challenges and New Benchmarks for Deep Learning in Robotic Vision*, June 2018.

[9] 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.

[10] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, 2001, pp. 131–140.

[11] D. Scharstein and R. Szeliski, "Stereo matching with nonlinear diffusion," *International journal of computer vision*, vol. 28, pp. 155–174, 1998.

[12] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun, "Continuous markov random fields for robust stereo estimation," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 45–58.

[13] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[14] M. Brown, D. Burschka, and G. Hager, "Advances in computational stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 993–1008, 2003.

[15] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 328–341, 2008.

[16] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox, "Flownet: Learning optical flow with convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2758–2766.

[17] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[18] Q. Wang, S. Shi, S. Zheng, K. Zhao, and X. Chu, "Fadnet: A fast and accurate network for disparity estimation," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 101–107.

[19] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[20] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *arXiv preprint arXiv:1709.00930*, 2017.

[21] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised learning of stereo matching," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1576–1584.

[22] F. Aleotti, F. Tosi, P. Z. Ramirez, M. Poggi, S. Salti, S. Mattoccia, and L. Di Stefano, "Neural disparity refinement for arbitrary resolution stereo," in *2021 International Conference on 3D Vision (3DV)*, 2021, pp. 207–217.

[23] Z. Yang, R. Simon, and C. Linte, "A disparity refinement framework for learning-based stereo matching methods in cross-domain setting for laparoscopic images," in *Proceedings of SPIE–the International Society for Optical Engineering*, vol. 12466. NIH Public Access, 2023.

[24] J. Zhang, Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Learning stereo matchability in disparity regression networks," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 1611–1618.

[25] K. Batsos and P. Mordohai, "Recresnet: A recurrent residual cnn architecture for disparity map enhancement," in *2018 International Conference on 3D Vision (3DV)*, 2018, pp. 238–247.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Z. Chen, W. Long, H. Yao, Y. Zhang, B. Wang, Y. Qin, and J. Wu, "Mocha-stereo: Motif channel attention network for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 768–27 777.

[29] P. Xu, Z. Xiang, C. Qiao, J. Fu, and T. Pu, "Adaptive multi-modal cross-entropy loss for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5135–5144.

[30] X. Wang, G. Xu, H. Jia, and X. Yang, "Selective-stereo: Adaptive frequency information selection for stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 19 701–19 710.

[31] M. Feng, J. Cheng, H. Jia, L. Liu, G. Xu, and X. Yang, "Mc-stereo: Multi-peak lookup and cascade search range for stereo matching," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 344–353.