IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 December 2023, accepted 5 January 2024, date of publication 10 January 2024,
date of current version 17 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3352136

RESEARCH ARTICLE

A Robust End-to-End Speckle Stereo Matching Network for Industrial Scenes

YUNXUAN LIU, KAI YANG, XINYU LI, ZIJIAN BAI, YINGYING WAN, AND LIMING XIE
School of Physical Science and Technology, Southwest Jiaotong University, Chengdu, Sichuan 610031, China

Corresponding author: Kai Yang (yangkai_swjtu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771409, and in part by the Science and Technology Project of Sichuan under Grant 2021YJ0080.

ABSTRACT The detection capability of deep learning-based stereo matching in industrial applications is inherently limited due to challenges posed by weak texture and inconsistent reflectance, making it difficult to accurately recover complex surface details. To achieve accurate measurements, this paper presents an end-to-end speckle stereo matching network that incorporates fringe, Gray code, and speckle projection patterns. The model is trained using a high-precision dataset consisting of thousands of pairs generated through binocular Gray code-assisted phase shifting. After establishing local correspondences between the left and right images using speckle patterns, the images are used as inputs to the network. The proposed network consists of two siamese 2D feature extraction networks. One network is dedicated to cost volume computation, while the other focuses on weight refinement feature extraction. The former network incorporates a lightweight module for extracting high-dimensional fusion features. These features are obtained from different dilation scales and randomly concatenated along the channel dimension. Patch convolution is utilized to effectively adapt to pixel features at various levels, reducing redundancy within the cost volume and improving the network's capacity to learn from ill-posed regions. Experimental results demonstrate that the proposed network achieves a significant improvement of approximately 10.7% in matching accuracy compared to state-of-the-art networks on public datasets. Furthermore, this method exhibits outstanding matching results when applied to diverse industrial scenarios. The reconstruction error for the radius of optical standard spheres is below 0.06-mm, which meets the demands of the majority of industrial applications.

INDEX TERMS Stereo matching, Gray code-assisted phase shifting, high robustness industrial imaging, group-wise volume, correction map.

I. INTRODUCTION

Optical 3D measurement is extensively employed in various fields, including material science [1] and biometrics [2], due to its advantages of high speed, high accuracy, and non-contact nature. Based on different illumination and imaging approaches, it can be categorized into passive and active measurements. Active measurement techniques offer higher reconstruction accuracy in textureless regions, making them particularly suitable for industrial inspection [3]. Obtaining comprehensive three-dimensional information for key components of trains is crucial for the holistic health monitoring of high-speed and subway rail systems. Commonly employed measurement techniques, include static measurement methods, light sectioning, and multi-line laser scanning. Nonetheless, these methods exhibit low measurement efficiency or fail to capture complete three-dimensional dimensions in a single attempt. Active 3D measurement, involving the use of an active light source for illumination, employs distinct modulation strategies based on the object's surface features to interpret the structured light field and subsequently derive the object's three-dimensional information. This approach finds applications in tasks such as wheel size measurement, bolt loss detection, automated guided vehicle

The associate editor coordinating the review of this manuscript and approving it for publication was Pinjia Zhang.

navigation, and robotic arm-assisted positioning. Nevertheless, the diverse dimensions of objects subjected to the holistic detection process for ensuring the secure operation of trains pose a challenge for traditional structured light methodologies in obtaining a comprehensive and dependable point cloud. Furthermore, the attainment of high-precision point clouds necessitates the aggregation of multiple frames, thereby constraining their applicability in dynamic measurement contexts.

Fringe and speckle patterns are two commonly used projection modes in structured light-based measurement methods. Speckle Profile Projection (SPP) utilizes speckle patterns projected with a spatial encoding strategy to provide local uniqueness for pixel labeling. However, ensuring the uniqueness of spatial pixels solely through projecting a single speckle pattern is challenging [4], [5]. To address the matching difficulties in SPP, local matching is often utilized, where the differences between pixels in each region are regressed for correction [6]. Alternatively, global matching estimates the disparity of all pixels directly by constructing an energy function that incorporates global information [7], [8]. However, these methods often trade off matching accuracy for reliable matches through disparity smoothing, making it challenging to obtain precise 3D information from a single-frame speckle projection.

In recent years, scholars have raised several deep learning-based stereo matching methods that exhibit higher accuracy and robustness compared to traditional algorithms [9]. Kendall employed a 3D-CNN to construct a 4D cost volume and utilized a soft attention mechanism, known as soft-argmin, to enable sub-pixel disparity regression [37]. To enhance the accuracy of feature extraction, Chang raised a method that converts local matching into a global stereo matching approach by utilizing patch convolution with different sizes [10], and the incorporation of a commonly used spatial pyramid pooling structure [40], by expanding the receptive field, it simultaneously removes the constraint of a fixed input image size. Numerous existing stereo matching methods predominantly focus on optimizing performance for specific datasets, resulting in limited generalization to other datasets. This limitation arises due to the susceptibility of these methods to domain shift, making it challenging to extend their performance to unexplored domains [11], [12]. For example, taking into account the occurrence of both positive and negative disparity in real-world scenarios, a semi-dense disparity map can be computed using binocular views, subsequently, the remaining regions can be completed using monocular views [13]. By utilizing pyramid-based warping cost volume, the fusion of multi-scale composite costs enables the extraction of domain-invariant features [14], the generalization from synthetic domains to real domains can be accomplished through the utilization of drone imagery and LIDAR point cloud reconstruction, enabling the generalization from the synthetic domain to the real domain. Recent research has demonstrated that by guiding and filtering the cost volume, it is possible to suppress redundant information, thereby simultaneously reducing the burden of cost aggregation and enhancing prediction accuracy. For instance, feature correlation can be effectively enhanced by employing image-guided weights [15]. Xu proposed a method that utilizes edge-preserving filtering with slice operations to effectively enhance the resolution of the cost volume [36]. In order to capitalize on the strengths of both the group-wise correlation volume [28] and the concatenated cost volume, Guo proposed a cost volume filtering technique that involves directly concatenating feature maps from different levels to compute the cost volume [43]. By utilizing stereo matching methods that employ edge-preserving filtering techniques [16], [17]. The contour information of the target object in the predicted results can be effectively enhanced, thus improving the preservation of its shape and edges, during the training process, accurate Ground truth disparity can be sparsely sampled by incorporating edge information and saliency information [18]. However, this approach may lead to inaccurate surface depth information. In addition to leveraging RGB images, utilizing non-visible spectral information has proven to be highly effective. One approach involves the use of an infrared projector and a single camera to construct a monocular infrared structured light system, which serves as guidance information [19]. However, this method has limitations in dynamic detection since it cannot achieve detection in a single-frame imaging manner. Nonetheless, three-dimensional reconstruction methods based on deep learning are often constrained by the uniqueness of the data. When the detected scene or target undergoes changes, the reconstruction accuracy of the model tends to degrade.

This paper proposed a single-frame stereo matching method designed for high-precision 3D measurements. Simultaneous acquisition of fringe patterns, Gray code patterns, and speckle patterns is performed for industrial scenes, a rich and high-precision dataset consisting of 6480 pairs of scenes was constructed using a combination of binocular Gray code and phase shifting techniques, the network takes images in speckle pattern as input [20]. In contrast to other frequently employed network architectures that utilize pyramid pooling structures for feature extraction, the network initially constructs a lightweight cascaded encoder-decoder module to extract high-dimensional fused features, the features acquired from various dilation scales are randomly concatenated along the channel dimension, resulting in enhanced matching capability for speckle points. Patch convolution is utilized to adapt to pixel features at various levels, enhancing the network's ability to refine the cost volume and suppress redundant information, this process further improves the feature matching capability of the network, the precision of the edge regions in the disparity map is strengthened by incorporating an additional edge loss function. Through a series of experiments, this method has been proven to achieve high precision and robustness in sub-pixel 3D reconstruction.

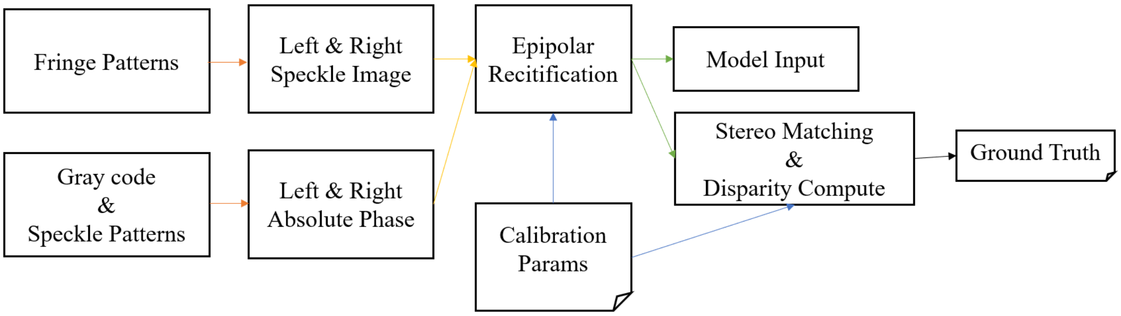

Fringe Patterns
Left & Right Speckle Image
Epipolar Recitification
Model Input
Gray code & Speckle Patterns
Left & Right Absolute Phase
Stereo Matching & Disparity Compute
Ground Truth
Calibration Params

FIGURE 1. Dataset production process, the figure illustrates the process of creating the dataset through three projection modes in this paper.

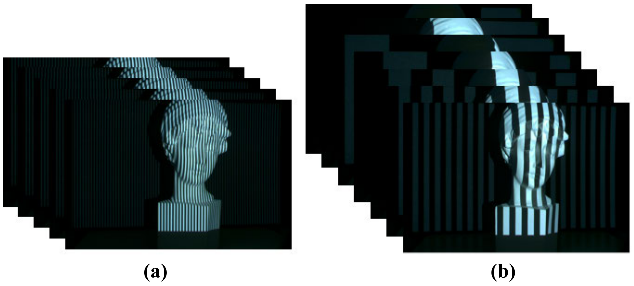


FIGURE 2. Gray code, fringe projection, the figure is from the 7 plus 5 Gray code phase shifting method used in the dataset generation.

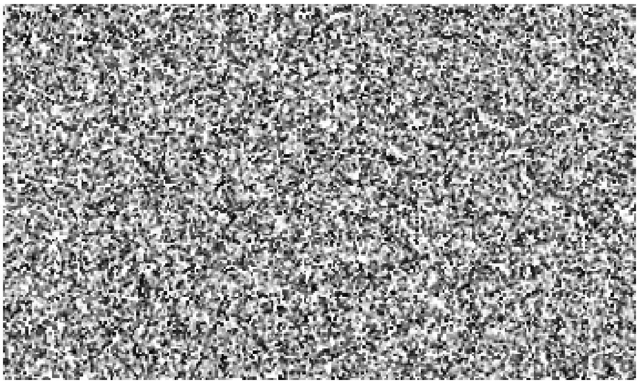


FIGURE 3. Example of Speckle Pattern, the pseudo-random speckle pattern generated using black and white binary structured pattern.

The remaining parts of this paper are as follows. Section II presents the proposed method, primarily encompassing the principles of the manufacture of training data and the design of the stereo matching network. In Section III, the passive measurement capability of the model is verified using public datasets. The performance of the proposed model in active measurement is evaluated using our speckle industrial dataset and optical standard balls. Furthermore, the feasibility of the overall approach proposed in this paper for industrial applications is discussed. Lastly, Section IV provides a summary of this paper.

II. METHOD

A. METHOD OF DATASET CONSTRUCTING

Fig.1. depicts the process of dataset creation in this study. It entails projecting Gray code patterns and stripe patterns to calculate the absolute phase of the left and right images. Following epipolar rectification, stereo matching and disparity calculation are performed to obtain the Ground truth. Speckle patterns are employed to establish spatial local connections, serving as inputs to the model.

The standard phase shifting profilometry (PSP) technique utilizes a set of sinusoidal fringes [21] that undergo equal phase shifts within one cycle and are projected onto the target scene, as shown in Fig.2(a).

The intensity distribution of the captured fringe patterns by the cameras is as follows:

$$I(x, y) = A(x, y) + B(x, y)\cos(\varphi(x, y) - \frac{2\pi n}{N}) \qquad (1)$$

where *A(x,y)* represents the background light intensity, *B(x,y)* represents the modulation degree, n represents the phase shift index with $n = 0, 1, 2, \ldots, N$-1, and $\varphi(x, y)$ represents the corresponding phase, which can be obtained [22] by the following:

$$\varphi(x, y) = \arctan \frac{\sum_{n=1}^{N-1} I_n(x, y)\sin(\frac{2\pi n}{N})}{\sum_{n=1}^{N-1} I_n(x, y)\cos(\frac{2\pi n}{N})} \qquad (2)$$

In (2), the wrapped phase is calculated using the arctangent function. In this case, the phase $\varphi(x, y)$ is truncated within the range of $(-\pi, \pi)$, and it is necessary to unwrap the phase to restore it to a continuous phase. Gray code utilizes a projection mode with black and white fringe [23], offering the advantages of high speed and error-free transmission, as shown in Fig.2(b). In this paper, it is utilized for phase unwrapping in phase shifting method [24]. By projecting N sets of Gray code images, 2^N periods of fringe patterns are marked, and unique identification can be recognized through the binary intensity sequence.

Before decoding the Gray code, it is essential to perform binary thresholding on the Gray code images. The threshold is determined based on the fringe pattern image:

$$A(n) = \frac{1}{m}\sum_{i=1}^{m} I_i(n) \qquad (3)$$

$$\Phi(x, y) = \varphi(x, y) + 2\pi k(x, y) \qquad (4)$$

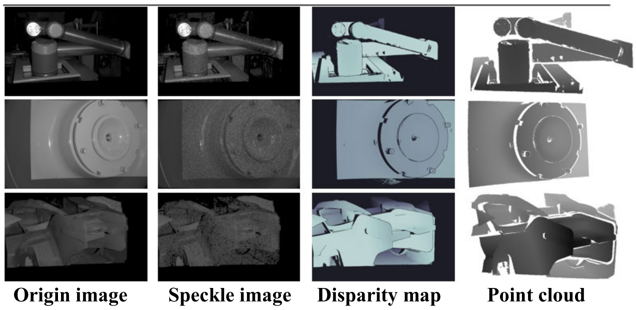


FIGURE 4. Schematic diagram of the proposed network structure, from left to right are: the image under normal illumination, the image under pseudo-random speckle projection mode, the Ground truth of the disparity map, and the point cloud Ground truth obtained using calibration parameters.

where m represents the number of captured fringe patterns, $I_i(n)$ denotes the grayscale value of a pixel during the projection of sine fringe patterns. The absolute phase $\Phi(x, y)$ of the left and right images can be represented by (4), where k represents the decoded Gray code level.

Performing epipolar rectification on the absolute phase. The disparity of the absolute phase between the left and right images is computed based on the principles of binocular imaging [25]:

$$D(x, y, Z) = \min\{|I_R(x, y) - I_L(x + d, y)|, disp\} \tag{5}$$

where D represents the disparity with respect to the left image as the reference, I_L and I_R denote the absolute phase pixel values corresponding to the left and right images, and *disp* represents the maximum estimated disparity in the current scene.

Obtaining three-dimensional information about a scene through passive single-frame stereo vision still faces several challenges. In industrial scenarios, many objects under inspection exhibit discontinuous surfaces, weak textures, or low reflectance. To enhance target features and improve matching accuracy in such cases, active illumination through controlled lighting is required. Laser speckle [26], [27], with its operational simplicity and cost-effectiveness, finds widespread application in active three-dimensional imaging in industrial settings. However, in measuring abrupt changes in surfaces, it is challenging to obtain dense disparity due to the discontinuity of the surface. Therefore, this study employs DLP-projected speckle images [34]. In comparison to laser speckle, it possesses superior local randomness, global uniqueness, and higher matching accuracy. Initially, a black image of size 1280 × 720 is created. This image is then divided into several regions. Within each region, a random selection of pixels is made as seed points. Region growing is performed on these seed points, taking into account the continuity of pixel space, until a complete speckle image is generated. To ensure local randomness, it is required that the number of pixels with values of 255 or 0 in the image regions occupies approximately 42-45% of various sliding windows. Fig.3 presents the speckle pattern utilized in this paper.

The deep learning-based stereo matching network exhibits severe domain shift issues with the dataset [41], indicating poor generalization to data from different scenes. The underlying cause for this is conventional datasets are tailored to specific fixed scenes, acquiring weights based on target colors and shapes. When there are changes in ambient lighting or when the model encounters previously unlearned target objects, it fails to correctly match corresponding points. In order to address the issue of poor generalization in the network due to domain-specific biases in the dataset, substituting local unique markers based on speckle patterns instead of conventional RGB information as inputs to the network model accurately resolves cross-domain generalization problems for different detection objects.

This paper adopts 7 Gray code images and 5 sine fringe images for binocular Gray code phase shifting imaging. Partial examples of the dataset are shown in Fig.4.

B. NETWORK ARCHITECTURE

In this paper, we propose an effective end-to-end speckle matching network, primarily designed to address the challenge of accurate 3D measurement in complex industrial scenes. We propose two pairs of siamese feature extraction networks: the cost feature extraction network and the cost weight extraction network. Firstly, in the cost feature extraction network, we incorporate a high-dimensional feature fusion module, which generates high-dimensional fused features during the process of down-sampling feature extraction from input images. This module takes as input the fused features at different dilation scales. After regression, the weight (Multip Weight) is generated to adjust the Group-wise [28] cost volume, resulting in the final cost volume.

As illustrated in Fig.5, the left and right speckle images are fed into two pairs of siamese networks. They enter the pyramid pooling module following the red arrows, and then the cost volume is calculated. Afterwards, a decoding-encoding and regression operation is conducted, resulting in a Weight Correction Map of size $1 \times H/4 \times W/4$. The obtained map is used to correct the Gwc-volume obtained through the blue arrow path. Subsequently, the volume is resized to $(disp_{\max} - disp_{\min}) \times H \times W$ through cost aggregation. Finally, the disparity regression and reprojection processes are applied to generate the point cloud model.

In the feature extraction, we incorporate a lightweight feature fusion module in the cost feature extraction section. This module utilizes depthwise separable convolutions, which decompose the conventional convolution operation into depthwise convolution and pointwise convolution. This approach ensures the same output while reducing computational complexity. Each Neck module is designed as follows: it consists of two convolutional layers with kernel size 1×1, with a 3×3 convolutional layer inserted in between. The purpose of these layers is twofold: one is to modify the number of channels, and the other is to downsample the tensor.

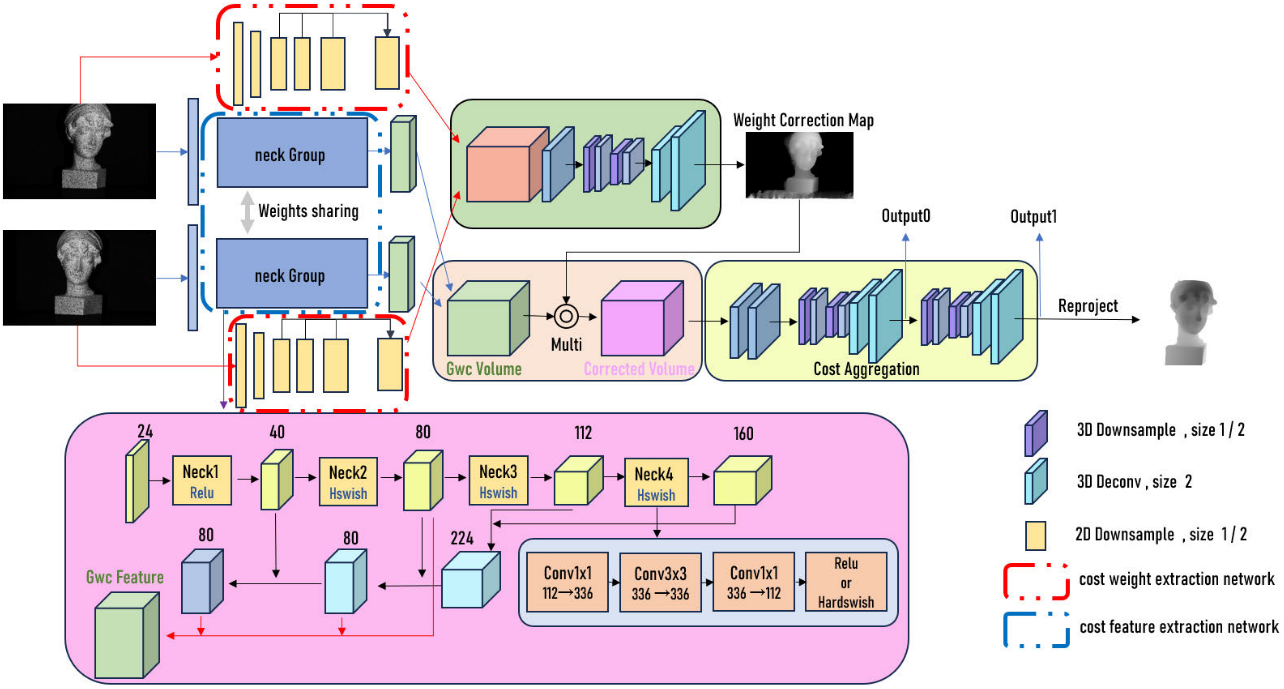


FIGURE 5. Schematic diagram of the proposed network structure,this network primarily consists of five components: feature extraction, 4D cost volume, cost volume correction, cost aggregation, and disparity regression.

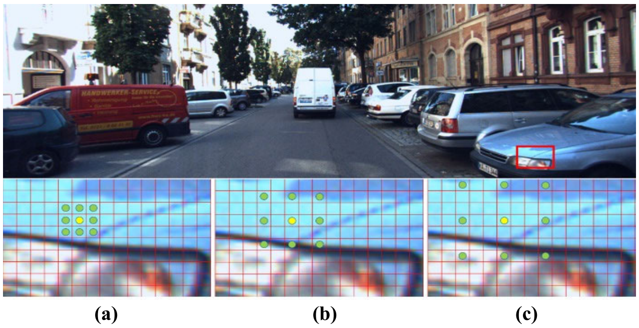


FIGURE 6. Pixel sampling weights for different expansion scales, (a), (b), and (c) respectively represent window modes of the same central point at different dilation scales.

Inspired by MobileNetV3 [29], we use a combination of h-swish and ReLU as activation functions in the cost feature extraction component. This choice strikes a good balance between computational speed and accuracy. Additionally, residual connections are utilized to enhance gradient propagation. To effectively utilize geometric information, features with different dilation scales are randomly concatenated as the output feature maps. As shown in Fig.6, this data is taken from the left image of the KITTI2015 test set. The second row of images highlights weakly textured regions within the red boxes. (a), (b), and (c) represent the pixel sampling weights for regular, dilation scale 2, and dilation scale 3, respectively, with the yellow color indicating the central pixel position. By representing the feature weights of the same pixel from different receptive fields, the network enhances the utilization of pixel neighborhood information. Fig.7 illustrates the differences between the proposed method and pyramid pooling feature extraction at the same level. The Color Bar indicates the magnitude of positive differences, with darker colors indicating larger differences. We observe that the proposed method shows a stronger feature representation in areas with limited texture.

Ultimately, the feature maps generated by this module have a final output size of 320 × $H/4$ × $W/4$. In the cost weight extraction network, we continue to utilize a pyramid-like structure [30] for weight extraction.

For the cost volume, this paper follows the approach of Group-wise correlation volume [28]:

$$c_{gwc} = \frac{N_p}{N_c}\left\langle m^p_{left}(x, y), m^p_{right}(x - d, y)\right\rangle \tag{6}$$

where N_C represents the number of channels in the 2D features, which are divided into N_P groups along the channel dimension. The "<, >" denotes inner product. This cost volume calculation method provides rich similarity features for 3D cost aggregation, reducing the parameter requirements. After obtaining the refinement weights (Multip Weight), the cost volume is further adjusted. The adjustment in the *k-th* channel of the cost volume follows:

$$C_{final}(k) = Mult(c_{gwc}(k), w_{Multip}) \tag{7}$$

where C_{final} represents the cost volume used for cost aggregation, *Mult* represents element-wise multiplication of matrices, and w_{Multip} represents the correction weight.

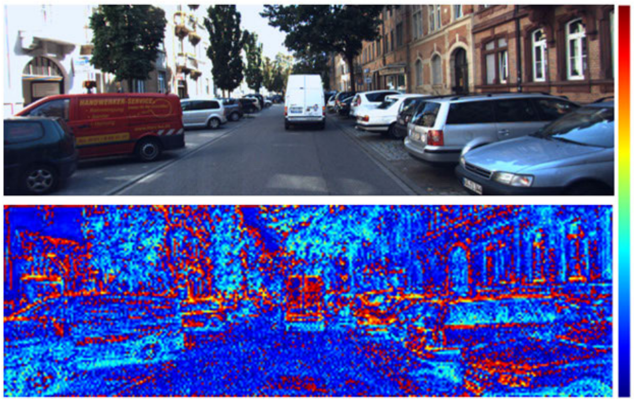


FIGURE 7. Feature map differences at the same level, the proposed method and the common pyramid pooling structure calculate the difference of feature map at the same layer, with brighter colors indicating larger differences in feature strength.

Cost aggregation aims to accurately reflect the correlation between pixels while aggregating feature information. Similar to previous 3D convolutional stereo matching networks, this paper utilizes stacked hourglass modules for cost aggregation. Each module consists of two three-dimensional convolutional layers with Batch normalization and ReLU activation, as well as two hourglasses module. During training, weighted losses are computed using the cost weight extraction network and the outputs of two decoders-encoders, and backpropagated to supervise the network. During testing, only the output of the second hourglass is used, and the disparity map is obtained by upsampling along the disparity dimension. The regression of predicted disparity values follows the soft-argmin mechanism using soft attention:

$$d = \sum_{D_{\min}}^{D_{\max}-1} l \cdot g_l \tag{8}$$

where l represents the disparity level, g_l represents the probability at that level, and D_{max} represents the maximum disparity trained by the model. Due to the limitations of Industrial scenes, there may be a disparity shift between the left and right fields of view, where the x-coordinate of a point in the left image is smaller than that in the right image. Therefore, a minimum negative disparity D_{min}, which needs to be learned.

To address the sharpening of target disparity edges, the predicted map and the Ground truth are first thresholded and dilated. The threshold for binarization is set to half of the normalized value of the predicted map. Then, the Binary Cross-Entropy loss is computed between the two to provide additional supervision for edge pixels:

$$L_{edge}(pre, Gt) = \sum_{k \in E} (Gt - 1)[\log(1 - pre)] - Gt[\log(pre)] \tag{9}$$

where pre and Gt represent the predicted results and the Ground truth.

The total loss function in this paper is defined as follows, where the Smooth L1 loss [37] is computed for the regression results:

$$\begin{aligned} L_{Total} = &\sum_{i=0}^{1} \lambda_i \cdot Smooth_{L1}(d_i - d_{gt}) \\ &+ \lambda_w \cdot Smooth_{L1}(d_w - d_{gt}) \\ &+ \lambda_E \cdot L_{edge}(pre, Gt) \end{aligned} \tag{10}$$

where d_i represents the predicted result of the *i-th* decoder-encoder, d_{gt} represents the Ground truth, E represents the spatial index of the edge pixels, and dw represents the result after disparity regression using cost weights, λ denotes the weight assigned to each individual loss.

III. EXPERIMENT

In this section, the capability of model in passive measurement is verified using public datasets Scene Flow [31] and KITTI [32], [33].The performance of the proposed model in active measurement is evaluated using our industrial dataset. The experimental datasets are described in Section III-A. Details of the experimental setup are presented in Section III-B. Evaluation metrics used in this paper and the effectiveness of the proposed modules and optimal settings are discussed in Sections III-C and III-D. Section III-E discusses the experimental results of our method on optical standard spheres. Section III-F analyzes the reconstruction results of our method in real industrial scenes and conducts a feasibility analysis.

A. DATASETS

Scene Flow datasets are synthetic stereo matching datasets consisting of three sub-datasets: Flyingthings3D, Monkaa, and Driving. It contains a total of 35,454 training data pairs and 4,370 testing data pairs. The images have a resolution of 960 × 540 and come with densely annotated Ground truth disparity.

KITTI2012 and **KITTI2015** are datasets for driving scenarios. KITTI2012 provides 194 training image pairs and 195 testing image pairs, with a resolution of 1226 × 370. KITTI2015 provides 200 training image sets and 200 testing image sets, with a resolution of 1242 × 375.

Our industrial datasets consist of two industrial cameras (Basler ace acA1920-40gc) with a resolution of 1920 × 1200 and a consumer-grade projector with a resolution of 1280 × 720. The fringe period is set to 16, and the working distance ranges from 0.5m to 1.5m. The camera baseline is 165mm, and the focal length is 12mm. In this study, we utilized 5 fringe patterns and 7 Gray code images.

The public datasets Sceneflow and KITTI are commonly employed for pre-training network weights and conducting network comparison experiments. They both utilize color (RGB information) as features to identify corresponding points for matching. However, due to limitations in industrial settings such as weak textures and non-continuous surfaces, this approach proves ineffective. Therefore, it is necessary to

TABLE 1. Model evaluation in different Settings.

Neck group	Loss edge	Multip Weight	KITTI 2015	Scene Flow	D1-All(%)	3px(%)	EPE(pixel)	D1(%)
			√	-	2.04	2.46		-
√	√		√	-	1.63	1.88		-
	√	√	√	-	1.75	1.96		-
√		√	√	-	1.49	1.56		-
√	√	√	√	-	**1.44**	**1.54**		-
			-	√	-		0.60	1.75
√	√		-	√	-		0.48	1.51
	√	√	-	√	-		0.52	1.62
√		√	-	√	-		0.48	1.45
√	√	√	-	√	-		**0.46**	**1.42**

TABLE 2. Evaluation of different methods (Scene Flow).

Method	Scene Flow	
	EPE(pixel)	D1 (%)
GcNet [37]	2.51	-
DispNetC [31]	1.68	-
CRL [38]	1.32	-
PsmNet [39]	1.09	3.89
GwcNet [28]	0.76	2.71
GaNet [40]	0.84	-
CfNet [41]	0.97	-
LEAStereo [42]	0.78	-
AcvNet [43]	0.48	1.59
Ours	**0.46**	**1.42**

locally annotate feature points by training the model using the poses projected by speckle patterns on the object surfaces. After obtaining qualified weights, achieving high-precision 3D reconstruction in industrial scenes only requires capturing a single frame with projected speckle patterns.

B. IMPLEMENTATION DETAILS

The model in this paper is implemented on the PyTorch framework and trained using the Adam optimizer [35] with $\beta 1 = 0.9$ and $\beta 2 = 0.999$. Network training and testing are performed on a Windows computer equipped with an Intel i9 9900X CPU and two NVIDIA TITAN RTX GPUs. The batch size of 6 is used during the training process, while the batch size of 4 is used during the testing process.

For the Scene Flow dataset, we used initial learning rate of 0.004, and the model is trained for 32 epochs. The learning rate is halved at the 6th, 16th, and 26th epochs. During training, the input data with the resolution of 960 × 540 is randomly cropped to 512 × 256.

For the KITTI dataset, this paper transfer learning from the Scene Flow dataset. The model is trained for 200 epochs with an initial learning rate of 0.0002, following a learning rate cosine annealing decay strategy (T_0 = 5, T_mult = 2).

Finally, the industrial scene dataset. The model is trained for 48 epochs, with a maximum disparity setting of 768 and a minimum disparity of -256.

C. ABLATION STUDY

In this section, we discuss the network performance under different settings, including the high-dimensional feature fusion module (Neck Group), the loss used for supervising edge and cost volume correction weight (Multip Weight). As shown in Table 1, the new modules significantly outperform the baseline(without any proposed models) setting. The Neck Group feature extraction module, edge loss and the Multi Weight structure reduce the three-pixel error (3PE) by 37.4% on the KITTI dataset and decrease the end point error (EPE) by 23.3% on the Scene Flow dataset. In this context, D1-All represents the disparity estimates that are considered as erroneous if they exceed the maximum value between 3PE and 0.05Gt, where Gt denotes the Ground truth disparity.

D. CONTRAST EXPERIMENTS

In this section, we conducted an evaluation and comparison of our proposed model with other models to validate its effectiveness. For the Scene Flow dataset, we utilized the EPE and D1. Fig.8 showcases the 3D reconstruction results obtained by PsmNet, BgNet, AcvNet, and our proposed method. For ease of visualization, zoomed-in results are displayed below the predicted results.

Our proposed method achieves more accurate disparity structures for each test sample. Compared to other methods, our approach demonstrates improved accuracy in areas with weak textures (image b,guitar fretboard) and small objects (image c,toy knife blade). This is attributed to the use of Multi Weight with different dilation scales in our network, which allows capturing cost volumes at various receptive field sizes. method. For ease of visualization, zoomed-in results are displayed below the predicted results.

Weight with different dilation scales in our network, which allows capturing cost volumes at various receptive field sizes.

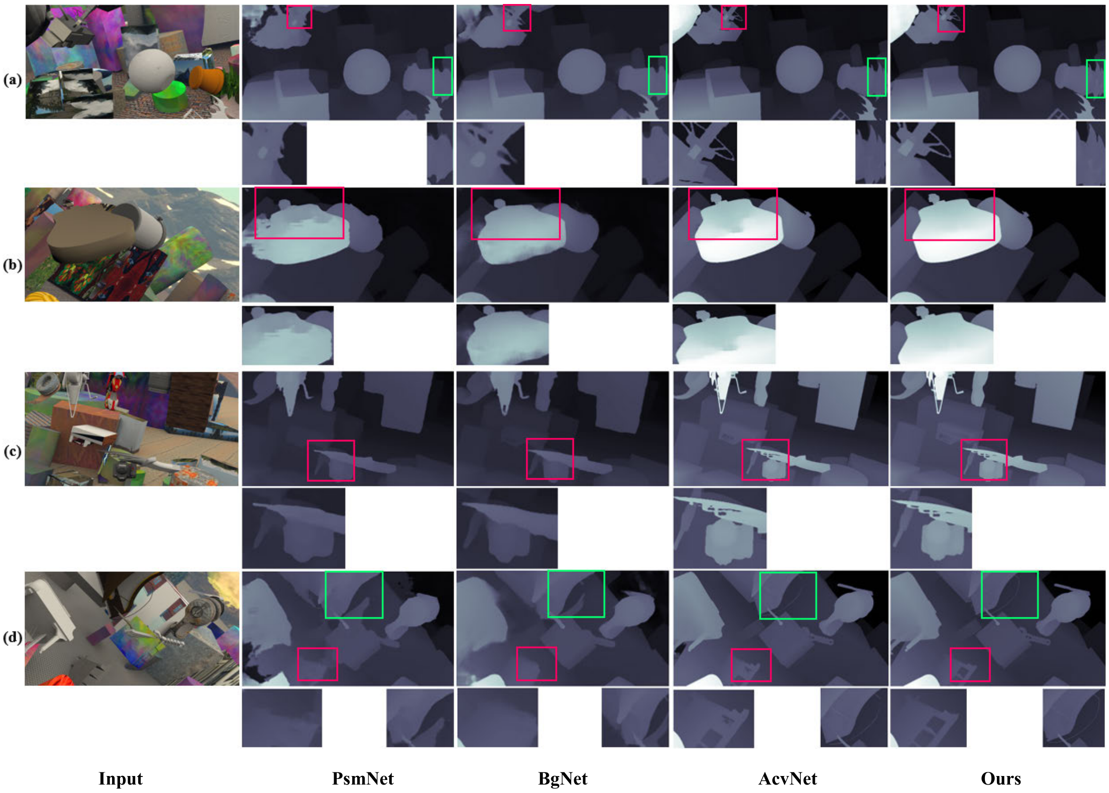


FIGURE 8. Comparison of methods (Scene Flow), from left to right, the images depict the original left image, PsmNet, BgNet, AcvNet, and the predicted results of our approach. The proposed method demonstrates decent accuracy in the pathological regions of the synthetic dataset.

TABLE 3. Evaluation of different methods (KITTI).

Method	KITTI 2012		KITTI 2015	
	2-All (%)	3-All(%)	D1-bg (%)	D1-All (%)
GcNet	3.46	2.30	2.21	2.87
PsmNet	3.01	1.89	1.86	2.32
EdgeStereo [44]	2.88	1.83	1.84	2.08
GwcNet	2.71	1.70	1.74	2.11
GaNet	2.50	1.60	1.48	1.81
AcfNet	2.35	1.54	1.51	1.89
HitNet [45]	2.65	1.89	1.74	1.98
CfNet	2.43	1.58	1.54	1.88
AcvNet	2.35	**1.47**	1.37	1.65
Ours	**2.12**	1.49	**1.31**	**1.44**

To quantitatively evaluate the performance of the compared methods, we compared nine different models and further calculated the average EPE and D1, as shown in Table 2. The dash (-) is used in table because these methods solely relied on the EPE metric for performance evaluation on the Scene Flow dataset, without considering the D1 metric. Overall, our method demonstrates a 10.7% improvement in D1 performance compared to the state-of-the-art AcvNet model in the testing results of the Scene Flow dataset.

To evaluate the effectiveness of our proposed model in complex scenes, we conducted comparative experiments on the KITTI2012 and 2015 urban street test datasets. Fig. 9 illustrates the disparity results comparison on the KITTI2015 dataset, with zoomed-in results displayed below the predicted results for better visualization.

Table 3 presents a performance comparison on the KITTI dataset. For KITTI2012, evaluation is conducted using the 3PE and the 2PE, where the maximum allowable error is set to three pixels (or two pixels). For KITTI2015, D1 errors are calculated for both background and foreground regions.

On the KITTI2012 dataset, our proposed method achieves a 9.8% improvement in 2PE compared to AcvNet, but experiences a slight retrogression of 1.36% in 3PE. However, for KITTI2015, our model demonstrates a significant performance improvement of 12.7%.

E. STANDARD SPHERE EXPERIMENT

To quantitatively analyze the accuracy of our proposed method, we fitted a sphere using 3D point cloud processing software to obtain the radius and center coordinates of the sphere. Fig.10 illustrates the three-dimensional information of optical standard spheres obtained in a non-laboratory setting. The output point cloud achieves dense and close-to-ground truth spherical point cloud. There are minor

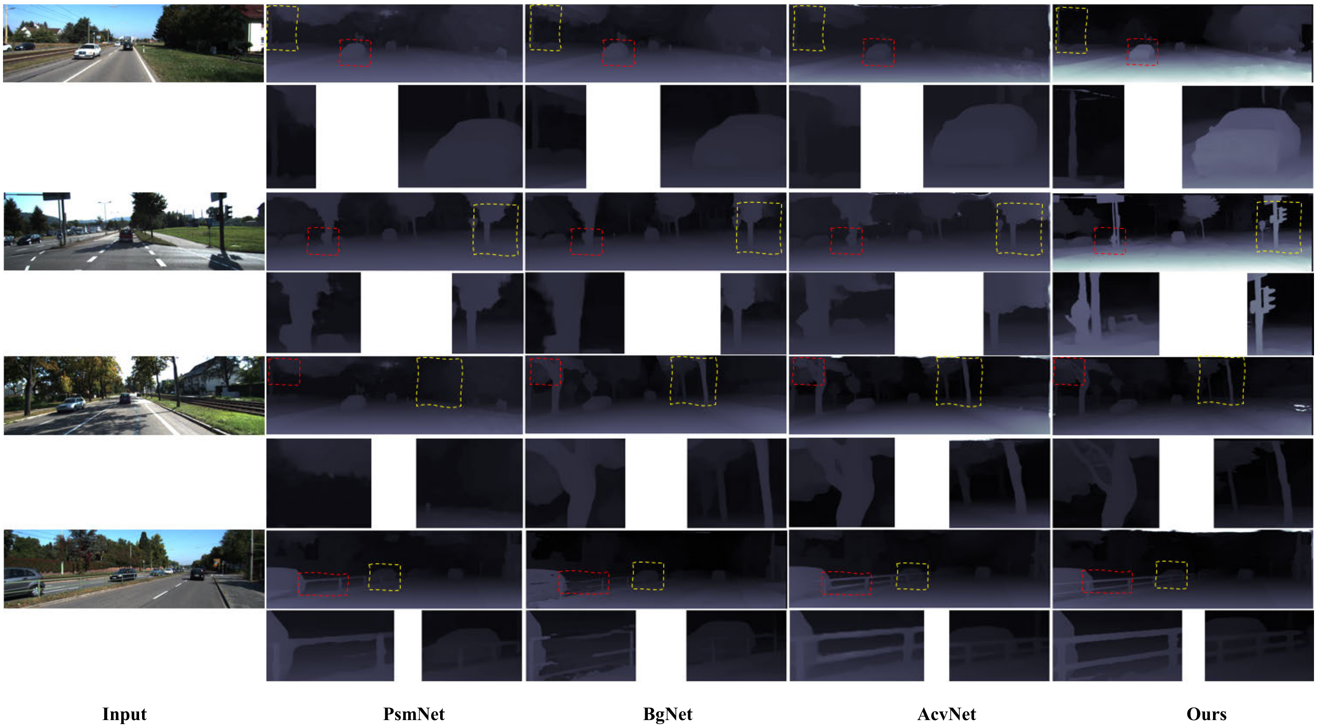


FIGURE 9. Comparison of methods (KITTI 2015), From left to right, the original left image, PsmNet, BgNet, AcvNet, and the predicted results of our method are shown. Our approach demonstrates excellent performance in reconstructing small objects.

TABLE 4. Comparison of standard sphere experiment.

	Radius (left)	RMS (left)	Radius (right)	RMS (right)	Distance (Center)
Standard	25.40mm	——	25.40mm	——	150.00mm
GT	25.43mm	0.1098	25.41mm	0.1359	150.48mm
prediction	25.45mm	0.1313	25.34mm	0.1408	151.26mm

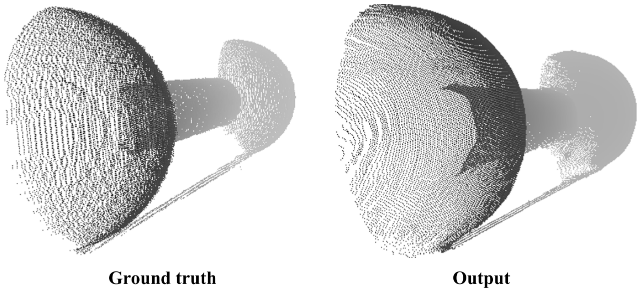


FIGURE 10. Standard sphere experience. The lateral view of the Ground truth and the output of the standard spherical point cloud.

inaccuracies present at the contact area between the sphere and the tabletop, which could be attributed to the difficulty of projecting speckle patterns accurately onto the contact region. Apart from that, there are no apparent protrusions or indentations in the remaining regions. Table 4 exhibits the parameters of the fitted sphere. It is worth noting that the standard sphere used as the test dataset achieves sub-pixel reconstruction accuracy, with a reconstruction radius error of no more than 0.06mm. This level of accuracy meets the precision requirements for industrial inspection.

F. INDUSTRIAL SCENARIO EXPERIMENTS

In this section, to thoroughly demonstrate the effectiveness of our proposed method, we trained the model using abundant training data. We conducted tests on common industrial scenes with weak textures and inconsistent reflectivity, such as wheels, snap fasteners and hollow axles.

Fig.11 presents some of the test results obtained from high-speed rail data in our experiments. Image *a-c* showcase the results for wheels, which contain large metallic areas. These areas are crucial and challenging to reconstruct in industrial scenes. The main components of the wheels consist of the tread surface and the axle. By using speckle images as input, our method successfully reconstructs dense disparity in regions where high reflectance phenomena. This helps mitigate disparity matching errors caused by non-uniform reflectivity to some extent. These results validate the robustness of our method in handling metal imaging scenarios effectively.

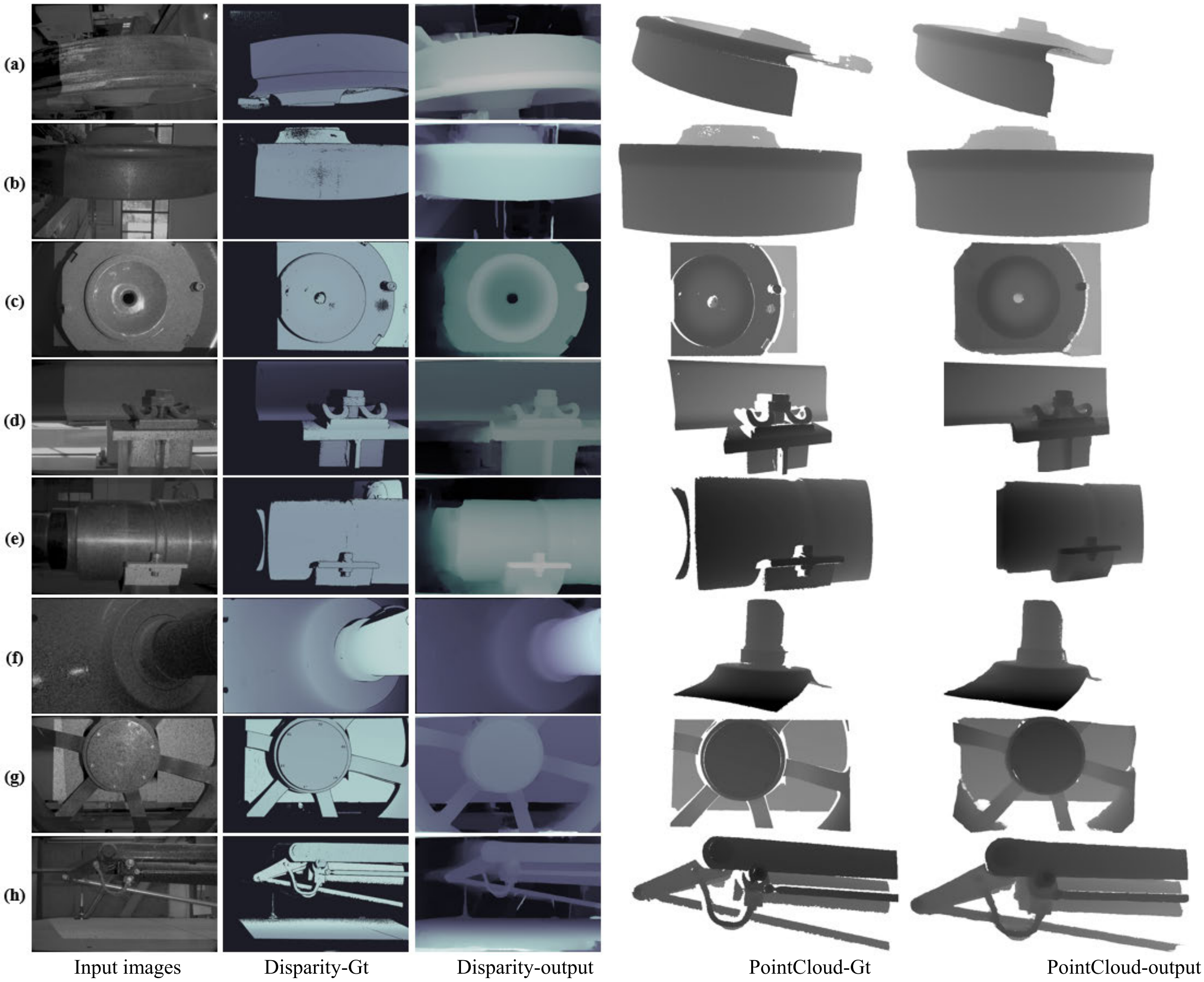


FIGURE 11. Test results on real railway data. These images are selected from a test dataset and all represent important components in the high-speed rail field. Similar to the standard ball, they are untrained data used to test the effectiveness and robustness of the proposed method.

Image *d* includes a steel rail with surface rust and a snap fastener. By establishing local correlations using speckle patterns, our proposed method is capable of mitigating erroneous matches caused by rusting.

For common reflective scenes, we performed 3D reconstruction on a hollow axle in Image *e*. The reconstruction was not affected by the large-area reflection along the axle's central axis, resulting in dense and continuous point clouds. Image *f* depicts the interface between the inner base of the wheel and the axle connection point. Approximately half of the field of view comprises low-texture regions, making it challenging to identify corresponding points. However, the proposed method in this paper enables clear reconstruction of the 3D information of the axle and base surface, which holds great significance for subsequent axle localization tasks.

Image *h* shows the pantograph pull rod on the roof of a train. Such scenes are typically challenging to reconstruct due to their special material properties. However, by utilizing a speckle projecting technique, our method can reconstruct dense and non-caking point clouds. This demonstrates the strong robustness of our proposed method in handling targets with low reflectance.

G. THEORETICAL ANALYSIS EXPERIMENTS

In order to facilitate a clear comparison between the conventional passive imaging method and the speckle imaging capability proposed in this study within an industrial setting, imaging experiments were conducted on the train bogie using both approaches.

The scene comprises the most challenging regions in stereo imaging, including low-texture and boundary areas, which are represented by the green and red boxes in Fig. 12.

Fig.13 presents the disparity distribution of conventional passive imaging and the proposed method in two regions. This validation approach has been commonly utilized in prior studies [37], [40].

As depicted in Fig.13(a), in the case of low-texture regions, the conventional method yields considerable noise as there

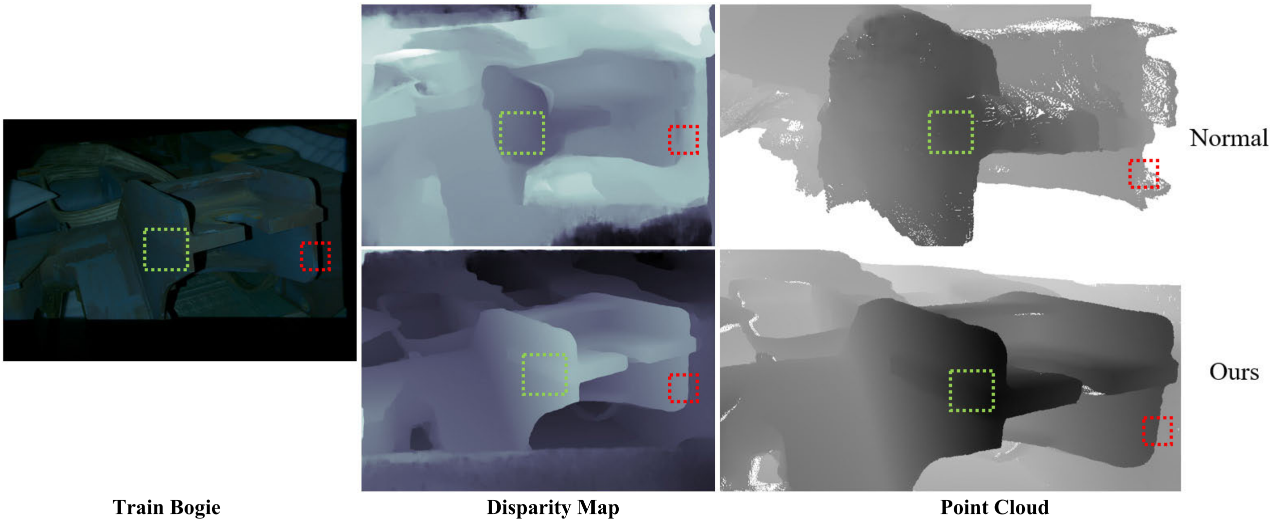
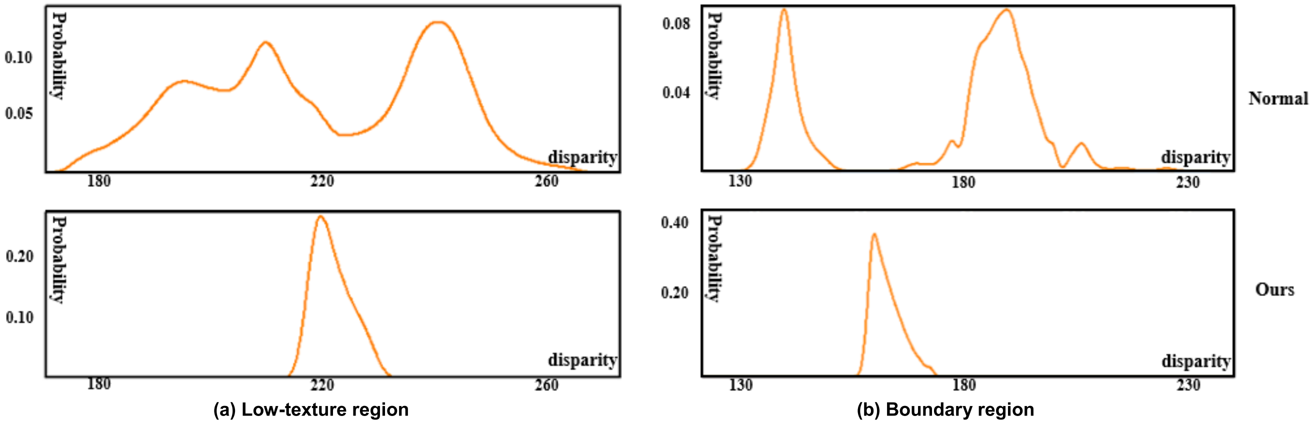


FIGURE 12. Reconstruction results of High-Speed train bogie. The first row of images represents the disparity map and point cloud results obtained through conventional passive imaging. The second row of images showcases the results obtained through the speckle projection method proposed in this study.



(a) Low-texture region

(b) Boundary region

FIGURE 13. Probability distribution of disparities after applying the soft-max. The samples were selected from the two most challenging regions: (a) Low-texture region, (b) Boundary region. The first row displays the results obtained through conventional passive imaging, while the second row showcases the results achieved using the method proposed in this study. The x-axis represents the disparity, while the y-axis represents the probability distribution of disparities.

are no prominent features available for accurate matching. The speckle projection pattern successfully suppresses noise in the disparity probability by establishing local feature uniqueness, leading to highly standardized and unimodal probability distributions.

The disparity discontinuity arising from the disparity jump between foreground and background leads to the presence of dual peaks in the disparity distribution of object edges. When disparity regression is further applied, it results in flocculent point cloud. As illustrated in Fig.13(b), the conventional method exhibits dual probability peaks influenced by both the background and foreground, resulting in fragmented point clouds. Conversely, the proposed method achieves accurate matching of the unimodal probability for edges, effectively eliminating the phenomenon of flocculent disparity along the edges.

In conclusion, when dealing with scenarios involving weak texture or non-contiguous regions in the target, persisting with passive measurement techniques, where the model learns object color or shape information, can lead to numerous erroneous matches. However, our proposed method, which involves locally labeling target features, effectively addresses the challenges of matching points in weak-textured regions and mitigates the issue of flocculent disparity caused by probability jumps.

IV. CONCLUSION

This paper proposes an end-to-end speckle stereo matching network that achieves high-precision, high-generalization, and highly robust sub-pixel 3D reconstruction. To ensure reliable network reconstruction, a dataset of 6,480 pairs high-precision stereo was created using 7 plus 5 binocular

Gray code phase shifting. Speckle images are used as network input. Quantitative analysis results demonstrate that the proposed network achieves approximately 10.7% higher matching accuracy compared to state-of-the-art networks on public datasets. It achieves radius reconstruction errors of no higher than 0.06mm for optical standard spheres and delivers excellent results on real high-speed rail data with weak texture levels and reflectivity non-uniformity.

There are several aspects of this method that could be improved. Firstly, due to the limited availability of GPU memory, a random cropping process is necessary during data input. While this allows for data augmentation and faster training, it can adversely affect the generalization of the learned weights to other data. Secondly, the proposed method demonstrates good performance in single-frame industrial inspections, but due to computational time constraints, it can only be applied to wheel-drop detection or low-speed rail-side inspections. Based on these analyses, we will explore lighter and faster stereo matching methods in the future.

REFERENCES

[1] J. Curt, M. Capaldo, F. Hild, and S. Roux, "An algorithm for structural health monitoring by digital image correlation: Proof of concept and case study," *Opt. Lasers Eng.*, vol. 151, Apr. 2022, Art. no. 106842, doi: 10.1016/j.optlaseng.2021.106842.

[2] T. Shi, Y. Qi, C. Zhu, Y. Tang, and B. Wu, "Three-dimensional microscopic image reconstruction based on structured light illumination," *Sensors*, vol. 21, no. 18, p. 6097, Sep. 2021, doi: 10.3390/s21186097.

[3] J. Zhang, W. Guo, Z. Wu, and Q. Zhang, "Three-dimensional shape measurement based on speckle-embedded fringe patterns and wrapped phase-to-height lookup table," *Opt. Rev.*, vol. 28, no. 2, pp. 227–238, Apr. 2021, doi: 10.1007/s10043-021-00653-9.

[4] M. Schaffer, M. Grosse, B. Harendt, and R. Kowarschik, "High-speed three-dimensional shape measurements of objects with laser speckles and acousto-optical deflection," *Opt. Lett.*, vol. 36, no. 16, p. 3097, Aug. 2011, doi: 10.1364/OL.36.003097.

[5] S. Heist, P. Dietrich, M. Landmann, P. Kühmstedt, G. Notni, and A. Tünnermann, "GOBO projection for 3D measurements at highest frame rates: A performance analysis," *Light, Sci. Appl.*, vol. 7, no. 1, p. 71, Oct. 2018, doi: 10.1038/s41377-018-0072-3.

[6] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *Int. J. Comput. Vis.*, vol. 33, no. 3, pp. 181–200, 1999.

[7] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 1068–1080, Jun. 2008.

[8] X. Zhang, H. Dai, H. Sun, and N. Zheng, "Algorithm and VLSI architecture co-design on efficient semi-global stereo matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4390–4403, Nov. 2020, doi: 10.1109/TCSVT.2019.2957275.

[9] Z. Liang, Y. Guo, Y. Feng, W. Chen, L. Qiao, L. Zhou, J. Zhang, and H. Liu, "Stereo matching using multi-level cost volume and multi-scale feature constancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 300–315, Jan. 2021, doi: 10.1109/TPAMI.2019.2928550.

[10] X. Ma, Z. Zhang, D. Wang, Y. Luo, and H. Yuan, "Adaptive deconvolution-based stereo matching net for local stereo matching," *Appl. Sci.*, vol. 12, no. 4, p. 2086, Feb. 2022, doi: 10.3390/app12042086.

[11] J. Zhang, X. Wang, X. Bai, C. Wang, L. Huang, Y. Chen, L. Gu, J. Zhou, T. Harada, and E. R. Hancock, "Revisiting domain generalized stereo matching networks from a feature consistency perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12991–13001, doi: 10.1109/CVPR52688.2022.01266.

[12] L. Kou, K. Yang, L. Luo, Y. Zhang, J. Li, Y. Wang, and L. Xie, "Binocular stereo matching of real scenes based on a convolutional neural network and computer graphics," *Opt. Exp.*, vol. 29, no. 17, p. 26876, Aug. 2021, doi: 10.1364/OE.433247.

[13] Y. Ding, Z. Li, D. Huang, Z. Li, and K. Zhang, "Enhancing multi-view stereo with contrastive matching and weighted focal loss," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 821–825, doi: 10.1109/ICIP46576.2022.9897772.

[14] Z. Shen, Y. Dai, X. Song, Z. Rao, D. Zhou, and L. Zhang, "PCW-Net: Pyramid combination and warping cost volume for stereo matching," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 280–297.

[15] A. Bangunharcana, J. W. Cho, S. Lee, I. S. Kweon, K.-S. Kim, and S. Kim, "Correlate-and-excite: Real-time stereo matching via guided cost volume excitation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 3542–3548.

[16] W. Xia, E. C. S. Chen, S. Pautler, and T. M. Peters, "A robust edge-preserving stereo matching method for laparoscopic images," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1651–1664, Jul. 2022, doi: 10.1109/TMI.2022.3147414.

[17] X. Cheng, Y. Zhao, W. Yang, Z. Hu, X. Yu, H. Zhao, and P. Zeng, "A novel cell structure-based disparity estimation for unsupervised stereo matching," *IET Image Process.*, vol. 16, no. 6, pp. 1678–1693, May 2022, doi: 10.1049/ipr2.12440.

[18] Q. Xu, S. Liu, G. Huang, K. Zeng, Y. Gong, and X. Luo, "CVE-Net: Cost volume enhanced network guided by sparse features for stereo matching," *Soft Comput.*, vol. 25, no. 24, pp. 15183–15199, Dec. 2021, doi: 10.1007/s00500-021-06257-4.

[19] Y. Xu, X. Yang, Y. Yu, W. Jia, Z. Chu, and Y. Guo, "Depth estimation by combining binocular stereo and monocular structured-light," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1736–1745.

[20] W. Yin, C. Zuo, S. Feng, T. Tao, and Q. Chen, "Learning-based absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation," *Proc. SPIE*, vol. 11571, pp. 149–158, Oct. 2020, doi: 10.1117/12.2573817.

[21] C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, "Phase shifting algorithms for fringe projection profilometry: A review," *Opt. Lasers Eng.*, vol. 109, pp. 23–59, Oct. 2018, doi: 10.1016/j.optlaseng.2018.04.019.

[22] V. Srinivasan, H. C. Liu, and M. Halioua, "Automated phase-measuring profilometry of 3-D diffuse objects," *Appl. Opt.*, vol. 23, no. 18, p. 3105, Sep. 1984, doi: 10.1364/AO.23.003105.

[23] Q. Zhang and Z. Wu, "Three-dimensional imaging technique based on gray-coded structured illumination," *Infr. Laser Eng.*, vol. 49, no. 3, 2020, Art. no. 303004, doi: 10.3788/IRLA202049.0303004.

[24] W.-C. Wang and W.-Y. Kang, "Measurement of surface topography of transparent objects by using digital phase-shifting shadow moiré method without painting," in *Advancement of Optical Methods in Experimental Mechanics*, vol. 3, H. Jin, C. Sciammarella, S. Yoshida, and L. Lamberti, Eds. Cham, Switzerland: Springer, 2014, pp. 221–227, doi: 10.1007/978-3-319-00768-7_28.

[25] C. Ma, Z. Sun, S. Pei, C. Liu, and F. Cui, "A road environment prediction system for intelligent vehicle," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–13, Apr. 2021, doi: 10.1155/2021/5569295.

[26] F. Salazar and A. Barrientos, "Surface roughness measurement on a wing aircraft by speckle correlation," *Sensors*, vol. 13, no. 9, pp. 11772–11781, Sep. 2013, doi: 10.3390/s130911772.

[27] R. Nothdurft and G. Yao, "Imaging obscured subsurface inhomogeneity using laser speckle," *Opt. Exp.*, vol. 13, no. 25, p. 10034, 2005, doi: 10.1364/OPEX.13.010034.

[28] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3268–3277, doi: 10.1109/CVPR.2019.00339.

[29] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[30] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[31] N. Mayer, E. Ilg, P. Häusser, 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048, doi: 10.1109/CVPR.2016.438.

[32] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361, doi: 10.1109/CVPR.2012.6248074.

[33] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3061–3070, doi: 10.1109/CVPR.2015.7298925.

[34] P. Zhou, J. Zhu, and H. Jing, "Optical 3-D surface reconstruction with color binary speckle pattern encoding," *Opt. Exp.*, vol. 26, no. 3, p. 3452, Feb. 2018, doi: 10.1364/OE.26.003452.

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

[36] B. Xu, Y. Xu, X. Yang, W. Jia, and Y. Guo, "Bilateral grid learning for stereo matching networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12492–12501.

[37] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 66–75.

[38] J. Pang, W. Sun, J. SJ. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 878–886.

[39] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5410–5418.

[40] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194, doi: 10.1109/CVPR.2019.00027.

[41] Z. Shen, Y. Dai, and Z. Rao, "CFNet: Cascade and fused cost volume for robust stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13901–13910.

[42] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 22158–22169.

[43] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12971–12980.

[44] X. Song, X. Zhao, H. Hu, and L. Fang, "EdgeStereo: A context integrated residual pyramid network for stereo matching," in *Computer Vision—ACCV 2018* (Lecture Notes in Computer Science), vol. 11365, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham, Switzerland: Springer, 2019, pp. 20–35, doi: 10.1007/978-3-030-20873-8_2.

[45] V. Tankovich, C. Häne, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "HITNet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14357–14367.

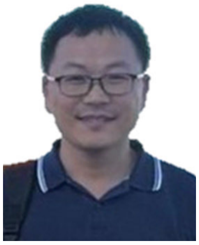
YUNXUAN LIU received the bachelor's degree in electronic information science and technology from the School of Physical Science and Technology, Southwest Jiaotong University, Sichuan, China, in 2021, where he is currently pursuing the master's degree in physics (optics). His research interests include stereo matching, 3D imaging, deep learning, and computer vision.

KAI YANG received the bachelor's, master's, and Ph.D. degrees from Southwest Jiaotong University, in 2003, 2006, and 2015, respectively. He is currently an Associate Professor with Southwest Jiaotong University. He is mainly engaged in the teaching of simulation program design and practice. His research interests include photoelectric detection and information processing, sensor technology, matlab simulation, digital image, and signal processing.

XINYU LI received the bachelor's degree in applied physics from the School of Physical Science and Technology, Southwest Jiaotong University, Sichuan, China, in 2021, where he is currently pursuing the master's degree in physics (optics). His current research interests include 3D imaging, computer vision, deep learning, and stereo matching.

ZIJIAN BAI received the master's degree in physics (optics) from the School of Physical Science and Technology, Southwest Jiaotong University. He worked at MEGVII Company, specializing in 3D imaging, deep learning, and computer vision. He is currently providing guidance with the School of Physics Science and Technology, Southwest Jiaotong University.

YINGYING WAN received the B.S. and Ph.D. degrees from Sichuan University, Chengdu, China, in 2012 and 2021, respectively. From 2018 to 2020, she was a Visiting Scholar with the University of Waterloo, Waterloo, ON, Canada. She is currently a Lecturer with Southwest Jiaotong University, Chengdu. Her research interests include 3D imaging and optical measurement.



LIMING XIE received the master's degree in physics from the School of Physical Science and Technology, Southwest Jiaotong University, and the Ph.D. degree, from the School of Physical Science and Technology, Southwest Jiaotong University, specializing in 3D measurement and optical imaging, computer vision, and deep learning.