Title:
3D Convolutional Feature Fusion for 3D Shape Reconstruction from Single-Frame Structured Light Image

Authors:
Jiyong Luo, ljyxox@stu.gxnu.edu.cn, Guangxi Normal University
Ming Chen, hustcm@hotmail.com, Guangxi Normal University
Shenglian Lu, lsl@gxnu.edu.cn, Guangxi Normal University

Keywords:
3D Shape Reconstruction, Depth Estimation, Fringe-Structured Light Image, Deep Learning

DOI: 10.14733/cadconfP.2025.116-120

Introduction:
The acquisition of three-dimensional information from objects holds significant importance for various fields such as robotic vision navigation, virtual reality (VR), industrial measurement, reverse engineering, and the demand for such information is steadily rising [1–4]. Standard methods for obtaining three-dimensional information from objects include binocular stereovision [5], time-of-flight [6], and structured light techniques [7] Structured light 3D measurement is a crucial way to obtain three dimensional information of an object. The principle is to project a specific encoded pattern onto the target object and then capture the image and decode the phase information of the encoded pattern. This phase of information reflects the difference in depth or height of the object's surface. By combining the phase information with the geometric relationship between the light source and the camera, the depth or height of each point on the surface of the object can be calculated.

For the depth estimation of structured light images, the current application of depth learning in structured light 3D measurement mainly focuses on using CNN to calculate the depth information from structured light images more accurately and faster, and realize the end-to-end process. Jeught et al. [9] proposed a CNN that can predict the 3D height of an input single-frame fringe-structured light image, which is the first end-to-end solution that uses a deep learning network to completely replace the phase demodulation process. Feng et al. [12] proposed a micro-depth learning contour measurement method, which can transform the input single-frame fringe-structured light image into the corresponding three-dimensional image. Nguyen et al. [13] proposed a robust method combining structured light technology and a deep convolutional neural network, which can predict the input of single-frame fringe structured light images and output corresponding depth maps. Jia et al. [15] proposed a new depth measurement method based on CNN, which can be regarded as a pixel-level classification regression task without matching, and depth information can be calculated from speckle structured light images without local stereo matching. Zhu et al. [8] combined the advantages of CNN and Transformer to design a two-branch network (CNN branch and Transformer branch). CNN branch and Transformer extract local features and global features from images, respectively. There are also some end-to-end solutions that use CNNS [16]. Although the aforementioned methods have achieved significant progress, depth prediction in regions such as small objects remains challenging. Moreover, these methods rarely focus on the crucial factor of receptive field. For dense prediction tasks, a larger receptive field can capture global contextual relationships, which helps improve prediction accuracy. Therefore, we propose a feature fusion method based on 3D convolution to expand the receptive field and capture global information from a single structured light image, thereby enhancing depth prediction accuracy. Extensive experiments on real-world datasets demonstrate the effectiveness of our method.

Compared to hNet [14] and UNet [18], our method improves accuracy by 17% and 16%, respectively, while using only 50% of their parameters.

Main Idea:

In neural networks, the receptive field refers to the size of the region of interest of a specific neuron in the feature map. Since any place outside the range of the receptive field in the input image does not affect the unit's value, it is necessary to control the receptive field size. In many visual tasks, especially dense prediction tasks such as semantic image segmentation, stereo vision, and depth estimation, large receptive fields can improve the accuracy of pixel-level localization and classification [17]. Therefore, we design a simple and effective method to expand the receptive field and use this method to design a 3D convolution feature fusion module to improve depth estimation accuracy. The 3D convolution feature fusion module is used to enlarge the receptive field. In addition, structural reparameterization can achieve a better trade-off between network accuracy and inference speed. In this section, we describe in detail the proposed method to expand the receptive field. Then, we use the method-based 3D convolution feature fusion module to design a simple and effective fringe structured light depth estimation network. The network core body uses an encoder-decoder, the encoder-decoder architecture can exploit global context information [10][11].The network structure is shown in Fig.1.

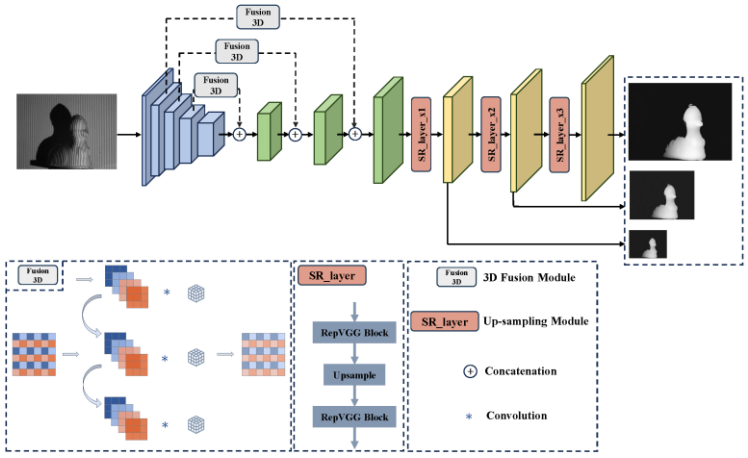


Fig. 1: Architecture overview of the proposed method.

Fusion 3D refers to the 3D convolution-based feature fusion module. The input to Fusion 3D is the feature map obtained during the feature extraction stage, and its output maintains the same resolution as the input. Fusion 3D first unfolds the input feature map f using a 2×2 window partitioning strategy, resulting in a feature map f_1 with four times the number of channels and half the spatial resolution. Then, f_1 is lifted into a higher-dimensional space to form f_2, which is processed by three consecutive 3D convolution operations for feature fusion in the high-dimensional space, producing f_3. Finally, a folding operation is applied to f_3 to restore it back to a 2D feature map f' with the original resolution.

The network first extracts multi-scale features from the input single-frame structured light image, generating feature maps $f_{1/4}$, $f_{1/8}$, and $f_{1/16}$ at 1/4, 1/8, and 1/16 of the input image resolution, respectively. The feature map $f_{1/16}$ is passed through the Fusion 3D module to obtain $f'_{1/16}$, which is then concatenated with f_0 to form *f_concat*. This concatenated feature is upsampled using 2D convolution to 1/8 of the input resolution. Similarly, $f_{1/8}$ and $f_{1/4}$ are progressively upsampled through the same strategy to produce a feature map f at 1/4 of the input resolution.

This feature map f is then fed into the second stage of the encoder. The employed *SR_Layer* module consists of two *RepVGG* Block [19] and a bilinear upsampling layer. *SR_Layer_1x* maintains the cur-

rent resolution, while *SR_Layer_2x* performs 2× upsampling. Intermediate supervision is applied to the outputs of all three SR_Layer modules, and the final depth prediction output is generated accordingly.

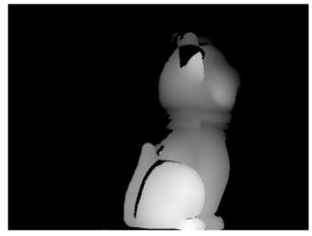
We conducted comparative experiments between our proposed method and state-of-the-art approaches on a real-world dataset. The dataset used is the open-source dataset proposed by Nguyen et al., which features structured light patterns with frequencies of 100, 20, 4, and 1. Depth maps are obtained using a four-frequency, four-phase multi-frequency heterodyne method. The experimental results are shown in Table 1.

Method	*Paramenters（M）*	*RMSE(mm)*	*Time(s)*
FCN[20]	-	2.03	-
AEN[13]	-	1.85	-
Unet[18]	8.63	1.62	**0.005**
hNet[14]	8.64	1.64	**0.005**
UNet-Wavele[21]	8.64	1.67	-
hNet-Wavelet[21]	8.64	1.59	-
DHDNet[16]	14.4	1.77	-
SIDO[22]	-	1.54	0.030
Our	4.53	**1.353**	0.010

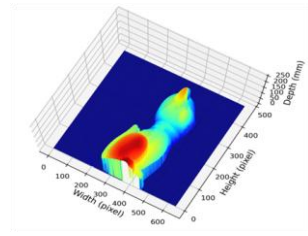
Tab. 1: Evaluation of the model on the test set of the real dataset.

As shown in Table 1, our proposed method achieves higher accuracy on the real-world dataset compared to state-of-the-art methods, while having the lowest number of parameters and only a slight increase in inference time. This demonstrates that our method achieves a good trade-off between speed and accuracy, highlighting the superiority of the proposed approach.

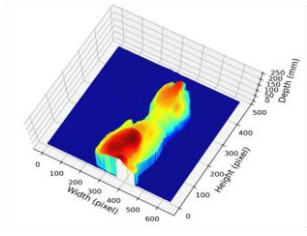
(a)Input fringe pattern



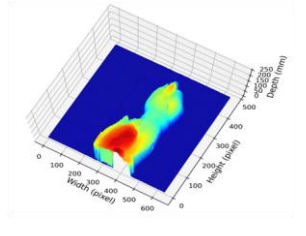
(b)Ground truth



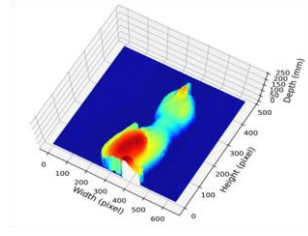
(c)Visualization of Ground truth



(d)Visualization of Unet's result



(e)Visualization of hNet's result



(f) Visualization of Ours result

Fig. 2: 3D visualization of model depth estimation.

As shown in Figure 2, our qualitative analysis also shows the 3D visualization results of UNet[18], hNet[14], and the depth estimation of the proposed model on the real dataset. On the real dataset, the

proposed model can obtain smoother and closer to the real depth results. At the same time, there are fewer anomalies. For example, as shown in Figure 6 (c) and Figure 6(d), the depth estimated by Unet and hNet models is abnormal in cat ears and cat tails, while the model proposed in this paper (i.e. (f) in Figure 1) does not.

Overall, our main contributions are as follows:

(1) We propose a method to enlarge the receptive field using 3D convolution and design a 3D convolution fusion module using this method.

(2) Based on the 3D convolutional fusion module, we design a simple and efficient structured light depth estimation network.

Conclusion:

In this paper, we design a network architecture for depth estimation from structured light images. The encoder-decoder structure is adopted in the core body of the network, and the structure reparameterization technology and 3D convolution feature fusion module are used. Structure reparameterization can achieve a favorable trade-off between the network's inference speed and accuracy, leading to high performance during the inference stage. The 3D convolutional feature fusion module can expand the receptive field. We propose that the network takes the fringe structured light image of a single frame as input and the output as the depth map of the corresponding image. We do complete experiments on the proposed network and other depth estimation networks on two datasets, and the experimental results show that our method is better than other methods in terms of parameter number and estimation accuracy, and can maintain a reasonable speed. In the region with rich details in the fringe-structured light image, the method in this paper is also better than other methods.

Acknowledgements:
We thank all the reviewers for their valuable comments. This research is supported by the Natural Science Foundation of Jilin Province (Grant No. 62062015, No. 61662006).

Jiyong Luo, https://orcid.org/0009-0005-0902-7677
Ming Chen, https://orcid.org/0000-0003-0506-5308
Shenglian Lu, https://orcid.org/0000-0002-4957-9418

References:

[1] Sansoni, G.; Trebeschi, M.; Docchio, F.: State-of-the-art and applications of 3d imaging sensors in industry, cultural heritage, medicine, and criminal investigation, Sensors, 9(1), 2009, 568–601. https://doi.org/10.3390/s90100568

[2] Li, B.; An, Y.; Cappelleri, D.; Xu, J.; Zhang, S.: High-accuracy, high-speed 3d structured light imaging techniques and potential applications to intelligent robotics, International journal of intelligent robotics and applications, 1(1), 2017, 86–103. https://doi.org/10.1007/s41315-016-0001-7

[3] Marrugo, A.G.; Gao, F.; Zhang, S.: State-of-the-art active optical techniques for three-dimensional surface metrology: a review, JOSA A, 37(9), 2020, 60–77. https://doi.org/10.1364/JOSAA.398644

[4] Zuo, C.; Feng, S.; Huang, L.; Tao, T.: Yin, W.; Chen, Q.: Phase shifting algorithms for fringe projection profilometry: A review, Optics and lasers in engineering, 109, 2018, 23–59, https://doi.org/10.1016/j.optlaseng.2018.04.019

[5] Dhond, U.R.; Aggarwal, J.K.: Structure from stereo-a review. IEEE Transactions on Systems, Man, and Cybernetics, 19(6), 1989, 1489–1510. https://doi.org/10. 1109/21.44067

[6] Wang, Z.: Review of real-time three-dimensional shape measurement techniques, Measurement, 156, 2020,107624. https://doi.org/10.1016/j.measurement.2020.107624

[7] Feng, S.; Chen, Q.; Gu, G.; Tao, T.; Zhang, L.; Hu, Y.; Yin, W.; Zuo, C.: Fringe pattern analysis using deep learning, Advanced Photonics, 1(2), 2019, 025001–025001. https://doi.org/10.1117/1.AP.1.2.025001

[8] Zhu, X.; Han, Z.; Zhang, Z.; Song, L.; Wang, H.; Guo, Q.: Pctnet: depth estimation from single structured light image with a parallel cnn-transformer network. Measurement Science and Technology, 34(8), 2023, 085402. https://doi.org/10.1088/1361-6501/acd136
[9] Jeught, S.; Dirckx, J.J.: Deep neural networks for single shot structured light profilometry, Optics express, 27(12), 2019, 17091–17101. https://doi.org/10.1364/OE.27.017091
[10] Chang, J.-R.; Chen, Y.-S.: Pyramid stereo matching network, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp.5410–5418. https://arxiv.org/abs/1803.08669
[11] Lin, X.; Sun, S.; Huang, W.; Sheng, B.; Li, P.; Feng, D.D.: Eapt: efficient attention pyramid transformer for image processing, IEEE Transactions on Multimedia, 2023, pp. 50-61. https://doi.org/10.1109/TMM.2021.3120873
[12] Feng, S.; Zuo, C.; Yin, W.; Gu, G.; Chen, Q.: Micro deep learning profilometry for high-speed 3d surface imaging. Optics and Lasers in Engineering 121, 2019, 416–427. https://doi.org/10.1016/j.optlaseng.2019.04.020
[13] Nguyen, H.; Wang, Y.; Wang, Z.: Single-shot 3d shape reconstruction using structured light and deep convolutional neural networks, Sensors, 20(13), 2020, 3718. https://doi.org/10.3390/s20133718
[14] Nguyen, H.; Ly, K.L.; Tran, T.; Wang, Y.; Wang, Z.: hnet: single-shot 3d shape reconstruction using structured light and h-shaped global guidance network, Results in optics, 4, 2021, 100104. https://doi.org/10.1016/j.rio.2021.100104
[15] Jia, T.; Liu, Y.; Yuan, X.; Li, W.; Chen, D.; Zhang, Y.: Depth measurement based on a convolutional neural network and structured light, Measurement science and technology, 33(2), 2021, 025202. https://doi.org/10.1088/1361-6501/ac329d
[16] Wang, L., Lu, D., Qiu, R., Tao, J.: 3d reconstruction from structured-light profilometry with dual-path hybrid network, Eurasip journal on advances in signal processing, 2022(1), 2022, 14. https://doi.org/10.1186/s13634-022-00848-5
[17] Luo, W.; Li, Y.; Urtasun, R.; Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks, Advances in neural information processing systems, 29, 2016.
[18] Ronneberger, O.; Fischer, P.; Brox, T.: U-net: Convolutional networks for biomedical image segmentation, Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, 2015, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
[19] Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J.: RepVGG: Making VGG-style convnets great again, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 13733–13742. https://doi.org/10.1109/CVPR46437.2021.01352
[20] Long, J.; Shelhamer, E.; Darrell, T.: Fully convolutional networks for semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
[21] Zhu, X.; Han, Z.; Song, L.; Wang, H.; Wu, Z.: Wavelet based deep learning for depth estimation from single fringe pattern of fringe projection profilometry, Optoelectronics Letters, 18(11), 2022, 699–704. https://doi.org/10.1007/s11801-022-2082-x
[22] Nguyen, A.-H.; Rees, O.; Wang, Z.: Learning-based 3d imaging from single structured-light image, Graphical Models, 126, 2023, 101171. https://doi.org/10.1016/j.gmod.2023.101171