www.cad-journal.net
Computer-Aided Design AND APPLICATIONS

3D Convolutional Feature Fusion for 3D Shape Reconstruction from a Single Frame Structured Light Image

Jiyong Luo[1], Ming Chen[2] and ShengLian Lu[3]

[1]School of Computer and Engineering, Guangxi Normal University, LjyXoX@stu.gxnu.edu.cn
[2]School of Computer and Engineering, Guangxi Normal University, hustcm@hotmail.com
[3] School of Computer and Engineering, Guangxi Normal University, lsl@gxnu.edu.cn

Corresponding author: Ming Chen, hustcm@hotmail.com

Abstract. Significant progress has been made in recent years in inferring 3D information from structured light images, mainly due to advances in deep neural networks. However, it remains challenging to accurately recover 3D details from regions with rich structures, and existing models often contain too many parameters, limiting their practical deployment. To address this, we propose a network based on an encoder-decoder framework, which combines structural reparameterization and a 3D convolutional feature fusion module. The reparameterization improves the balance between inference speed and accuracy, while the 3D fusion module expands the receptive field, which is critical for depth estimation. We conduct extensive experiments on public and custom datasets using conventional networks such as hNet and U-Net. Compared to these baselines, our model achieves better accuracy and efficiency while requiring only about half the parameters.

Keywords: 3D shape reconstruction, depth estimation, fringe structured light image, deep learning
DOI: https://doi.org/10.14733/cadaps.2026.326-339

1 INTRODUCTION

The acquisition of three-dimensional information from objects is of significant importance across various fields, including robotic vision navigation, virtual reality (VR), industrial measurement, and reverse engineering, with a steadily rising demand for such information [29],[14],[20],[38]. Standard methods for obtaining three-dimensional information from objects include binocular stereovision [4], time-of-flight [32], and structured light techniques [7]. Structured light 3D measurement is a crucial method for obtaining the three-dimensional information of an object. The principle involves projecting a specific encoded pattern onto the target object, capturing the image, and decoding the phase information of the encoded pattern. This phase information reflects the differences in depth or height across 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 object's surface can be computed, enabling the construction of a three-dimensional shape model of the object. Structured light 3D measurement technology offers the advantages of high precision, high resolution, and low cost for 3D reconstruction, making it a vital 3D imaging solution in both academic and industrial fields [29]. The classical structured light technology may involve multiple fringe images, complex algorithms, and intensive computation to determine the phase distribution, parallax, and depth, resulting in higher accuracy and slower reconstruction speed. With the rapid advancement of artificial intelligence, particularly deep learning, significant achievements have been made in the fields of computer vision [13], natural language processing [3], speech recognition [9], and even Domestic waste classification [15]. Deep neural networks, which are hierarchical in structure, consist of many layers, with each layer ' s output serving as the input to the next. The greater the number of layers, the more complex features the network can learn, and consequently, the larger the number of parameters required by the entire network. Deep learning has gradually been applied to structured light 3D measurement and depth estimation in recent years. Although deep learning methods have made significant strides in structured light 3D measurement, challenges remain in restoring global features and fine details of the scene from the projected images. There is still substantial room for improvement in achieving more accurate 3D information inference [36].

In this paper, we propose a simple and efficient fringe structured light depth estimation network for calculating depth information from a single frame of fringe structured light images. The core of the network is based on an encoder-decoder architecture, which can leverage global context information [2][17]. The 3D convolution feature fusion module is employed to expand the receptive field. Furthermore, structural re-parameterization facilitates a better trade-off between network accuracy and inference speed. Overall, the main contributions of this work 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.

2 RELATED WORK

In structured light 3D reconstruction, phase demodulation is a crucial step for extracting phase information from captured coded pattern images. In the early stages of applying deep learning to structured light 3D reconstruction, researchers attempted to use it for phase demodulation. In [7]Feng et al. used CNN to analyze fringe-structured light images. They first demonstrated that deep neural networks could be used to perform analysis of fringe-structured light. Shi et al. [30] further demonstrated the effectiveness of label enhancement and patch-based strategies in phase retrieval using deep learning. Zhou et al. [34] used a deep neural network to realize the automation of Fourier transform profilometry frequency selection. Yin et al. [33] used deep learning along with two sets of phase-shifted fringe images to eliminate phase blur during phase unwrapping, demonstrating that issues in this process can be addressed through deep learning. The above work initially applies depth learning to structured light image analysis. However, these methods only replace a subroutine in the phase demodulation algorithm and do not achieve an end-to-end pipeline from structured light images to 3D reconstruction.

For depth estimation from structured light images, current applications of deep learning in structured light 3D measurement mainly focus 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. [10] 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. [8] proposed a micro-depth learning contour measurement method that can transform a single-frame fringe-structured light image into its corresponding three-dimensional image. Nguyen et al. [25] proposed a robust method combining structured light technology and a deep convolutional neural network, which can

take single-frame fringe structured light images as input and output corresponding depth maps. Nguyen et al. [24] proposed an end-to-end network that can transform a single frame of speckle-structured light pattern images into corresponding point clouds. Then Nguyen et al. [23] introduced an H-shaped global guidance network path and multi-scale feature fusion into CNN and proposed hNet to further improve the accuracy of deriving 3D information from structured light images. Jia et al. [11] 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. [36] combined the advantages of CNN and Transformer to design a two-branch network (CNN branch and Transformer branch). The CNN branch and Transformer branch extract local and global features from the images, respectively. A bidirectional feature fusion module and a cross-feature multiscale fusion module are designed to integrate the local features and global features extracted from the two branches in order to achieve better depth estimation. Other end-to-end solutions based solely on CNNs have also been proposed [31].

Although previous methods have made significant progress, depth prediction remains challenging in regions containing small objects. Moreover, these approaches seldom emphasize the importance of the receptive field, which plays a critical role in dense prediction tasks. A larger receptive field can better capture global contextual relationships, thereby improving prediction accuracy.

To address these limitations, we propose a feature fusion method based on 3D convolution to expand the receptive field and extract global information from a single structured light image, leading to more accurate depth estimation. Inspired by previous works [25], [11], we adopt an encoder-decoder architecture as the core architecture of our network. To balance speed and accuracy, we incorporate structural re-parameterization techniques into the network design.

3 METHODS

In neural networks, the receptive field refers to the spatial extent of the input that a specific neuron in the feature map is sensitive to. Since regions outside the receptive field do not influence a neuron's output, controlling the receptive field size is essential.

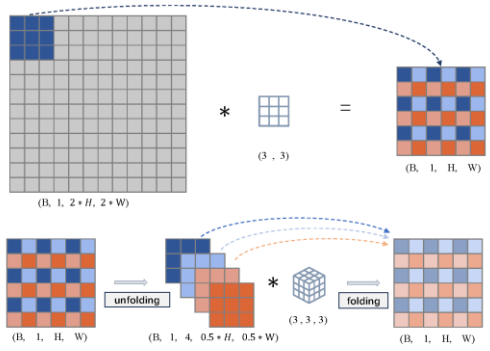


Figure 1: The method of using 3D convolution to enlarge the receptive field.

In many vision tasks, particularly dense prediction tasks such as semantic segmentation, stereo vision, and depth estimation, larger receptive fields can enhance the accuracy of pixel-level localization and classification [19],[16]. Therefore, we propose a simple and effective approach to expand the receptive field, which is then employed to construct a 3D convolutional feature fusion module for enhancing depth estimation accuracy. In this section, we provide a detailed description of the proposed receptive field expansion method. Subsequently, a simple and effective fringe-

structured light depth estimation network is constructed based on the proposed 3D convolutional feature fusion module.

3.1 The Method of Expanding the Receptive Field Based on 3D Convolution

As mentioned above, a large receptive field is essential for accurate depth estimation, especially in structured light scenarios where global context plays a key role in disambiguating local patterns. To this end, we introduce a receptive field expansion strategy based on 3D convolution, as illustrated in Figure 1.

Our approach is inspired by the feature unfolding mechanism in MobileViT [21], which was originally designed to convert 2D features into a format suitable for Transformer processing. We extend this idea to transform features along ascending and descending spatial dimensions, enabling efficient integration of 3D contextual information using convolutional operations. Compared to traditional approaches such as dilated convolutions or vanilla attention, this strategy offers a better balance between receptive field size and computational cost, making it well-suited for real-time structured light applications.

The transformation process consists of the following steps:

1)Set the split window size to (w_h,w_w).

2)According to the size of the window, the original feature map with the size of (b,c,(h$\times w_h$),(w$\times w_w$)) blocks is divided into (b,(c$\times w_h \times w_w$),h,w), the number of channels of the feature map becomes multiple of the original $w_h \times w_w$, and the height and width of the feature map are changed to the original $1/w_h$,$1/w_w$.

3)Finally, a new dimension is added to the new feature map, and its size is changed (b,1,(h$\times w_h$),(w$\times w_w$)), at this time,3D convolution can be used for feature map. Where b represents the batch size, c is the number of channels, h and w are the height and width of the feature map.

Here, b denotes the batch size, c is the number of channels, and h, w represent the height and width of the feature map, respectively. This transformation enables our model to benefit from volumetric feature aggregation with minimal overhead

3.2 Network Structure

Based on our method of using 3D convolution to enlarge the receptive field, we design a simple and effective fringe structured light depth estimation network. The network input is a single frame fringe pattern, and the output is the corresponding depth map. The core body of the network adopts an encoder-decoder structure, which includes a feature extraction layer based on structural re-parameterization, a fusion module based on 3D convolution to expand the receptive field, and a lightweight super-resolution up-sampling module. The network structure is shown in Figure 2.

3.2.1 3D feature fusion module

To effectively capture long-range information and leverage the advantages of fringe patterns, we design a 3D convolutional fusion module based on the receptive field expansion method described in Section 3.1. The module takes a 2D feature map as input and outputs a 2D feature map of the same spatial size. First, the input feature map f is unfolded using a window size of 2×2, resulting in a new feature map f_1 with four times the channel dimension and half the height and width. Then, an additional dimension is added to f_1 to obtain f_2, which is subsequently passed through three consecutive 3D convolution operations to perform feature fusion along the expanded dimension, yielding f_3. Finally, f_3 is folded back to restore the original 2D format, producing the output feature map f'.

3.2.2 Structural re-parameterization technology

To balance inference speed and accuracy, we adopt structural re-parameterization to design the network's encoder and decoder. This technique was originally introduced in RepVGG [5]. enables the construction of a simple yet efficient convolutional neural network architecture that achieves

high performance during inference. The core idea is to use a multi-branch convolutional structure during training, which is then merged into a single-path convolutional block at inference time.

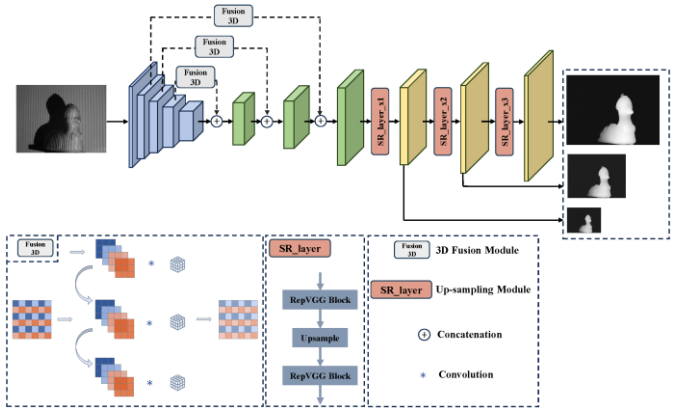


Figure. 2: Architecture overview of the proposed 3DFNet.

Each convolutional block comprises three components: a 1×1 convolution, a 3×3 convolution, and an identity mapping. This results in a re-parameterized network that combines low computational complexity and parameter count with strong representational capacity. The corresponding core formulation is as follows:

$$\begin{aligned} M^{(2)} = & bn(M^{(1)} * W^{(3)}, \mu^{(3)}, \sigma^{(3)}, \gamma^{(3)}, \beta^{(3)}) \\ & + bn(M^{(1)} * W^{(1)}, \mu^{(1)}, \sigma^{(1)}, \gamma^{(1)}, \beta^{(1)}) \\ & + bn(M^{(1)} * W^{(0)}, \mu^{(0)}, \sigma^{(0)}, \gamma^{(0)}, \beta^{(0)}) \end{aligned} \tag{1}$$

Where $M^{(1)}$ and $M^{(2)}$ denote the input and output, respectively. W,μ,σ,γ and β represent the convolutional kernel weights, the accumulated mean of the Batch Normalization (BN) layer, the standard deviation of the BN layer, the learnable scaling factor of the BN layer, and the bias term, respectively. The superscripts indicate the corresponding convolution type:3×3,1×1,and 0×0(identity mapping).

In this work, RepVGG [5] block is used as the fundamental convolution structure. The network architecture for both training and inference under the structural re-parameterization technique is illustrated in Figure 3.

The encoder part of this network follows a VGG-style framework, consisting of five stages in total. In the first stage, a single RepVGG [5] block is used to downsample the input single-channel fringe image to half of its original size, while the number of channels is increased to 32. In the second, third, and fourth stages, 2, 4, and 14 RepVGG blocks are applied to reduce the feature map size to 1/4, 1/8, and 1/16 of the original, respectively, with the number of channels set to 32, 64, and 128. In the fifth stage, a RepVGG block is used to increase the number of channels to 256, maintaining the resolution of the input feature map. Finally, the features from the second to fifth stages, denoted as $f_{1/4}$, $f_{1/8}$, $f_{1/16}$, and f_0, form the final output of the encoder.

3.2.3 Lightweight super-resolution up-sampling module

In convolutional neural networks, the decoder functions to transform the features extracted by the encoder into the target output. In this section, the decoder progressively upsamples the feature maps to the original input resolution and generates the corresponding depth map for the fringe image. The proposed decoder architecture consists of two main components.

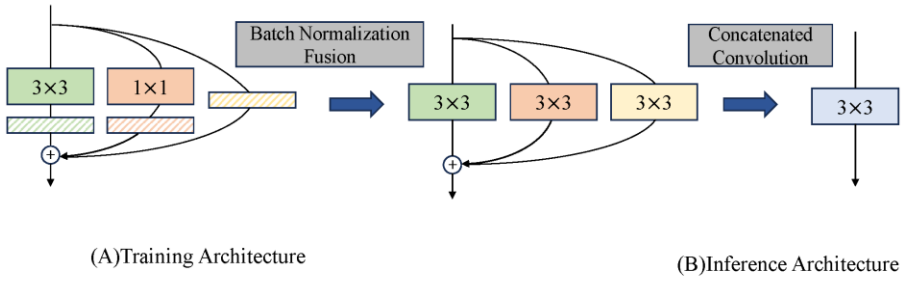


Figure 3: RepVGG [5] block structure.

First, for the encoder outputs $f_{1/4}$, $f_{1/8}$ and $f_{1/16}$, their corresponding enhanced features with long-range contextual information, denoted as $f'_{1/4}$, $f'_{1/8}$ and $f'_{1/16}$, are obtained using the 3D convolution fusion module. Then, $f'_{1/16}$ is concatenated with f_0, and the fused feature is upsampled to 1/8 of the original resolution via 2D convolution and bilinear interpolation. The remaining features are likewise progressively upsampled to 1/4 of the original resolution.

Second, a lightweight super-resolution module, denoted as the SR-Layer, is employed. It consists of a RepVGG block, a bilinear upsampling module, and another RepVGG block, as illustrated in Figure 2. Where SR-Layer-1x indicates no upsampling, while SR-Layer-2x denotes double upsampling. Intermediate supervision is applied to the outputs of the three SR-Layer modules, and their respective losses are calculated during training.

3.2.4 Loss function

The loss function plays a critical role in network training. In this work, we adopt the Root Mean Square Error (RMSE) loss to guide the optimization process, which is defined as:

$$Loss = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{Z}_i - Z_i)^2} \tag{2}$$

Where $\hat{Z}_i$ and Z_i denotes the predicted depth and ground-truth depth of the i-th point in the image, respectively, and N is the number of valid points.

4 EXPERIMENT

In this section, we detail the training procedure of our model, compare its performance with existing methods on a public dataset, and conduct ablation studies to demonstrate the effectiveness of the proposed approach.

4.1 Dataset and Evaluation Criteria

Real-world dataset: This paper uses the open-source real-world dataset proposed by Nguyen et al. [25], which contains 648 fringe structured light samples and 648 speckle structured light samples. We use the fringe structured light samples to train our network. Among them, 90% are used for training and validation, while the remaining 10% are reserved as an independent test set. As shown in Figure 4, the depth maps in this dataset are dense and highly accurate.

Synthetic dataset: Currently, there is no publicly available synthetic structured light dataset. Therefore, we follow the approach proposed by [26] and generate a synthetic structured light dataset using Blender [1]. In recent years, Blender has gained traction in academic research; for instance, it has been used for 3D visualization of biological macromolecules and astronomical data

[6][12], as well as for generating semantically labeled point cloud datasets and light field structured light projection data [27][37] Thus, using Blender to create a fringe structured light dataset is both feasible and scientifically sound. The synthetic dataset we created is easily scalable, and the diversity of samples can be enhanced by increasing the number of 3D models and applying various material textures. With minimal modifications, this dataset can be adapted for depth estimation tasks involving monocular, binocular fringe, and speckle structured light. In this study, we focus on single-frame fringe structured light depth estimation. The dataset consists of 12,600 samples, which are divided in a 6:2:2 ratio for training, validation, and final testing.

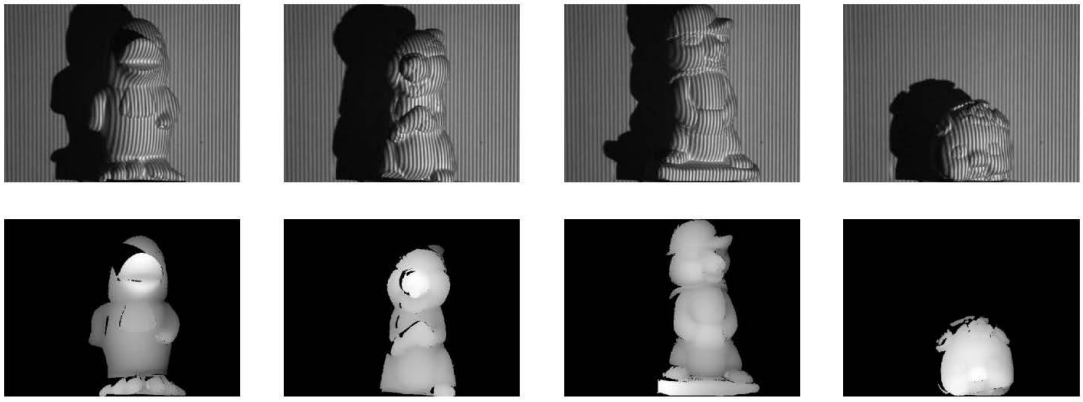


Figure 4: The first row shows the real-world fringe dataset, and the second row shows the corresponding depth map.

Evaluation criteria: To evaluate the effectiveness of the proposed model, we adopt commonly used evaluation metrics for fringe structured light depth estimation tasks, including RMSE, MSE, and PSNR. Among them, RMSE is the most widely used indicator. These metrics are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{z}_i - z_i)^2} \tag{3}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{z}_i - z_i)^2 \tag{4}$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{p_{\max}^2}{MSE}\right) \tag{5}$$

Where $\hat{z}_i$ and z_i denote the predicted depth and ground-truth depth of the i-th point in the image, respectively, and N is the number of valid points.

4.2 Implementation Details

The proposed model was implemented in PyTorch and trained on a workstation equipped with an Intel® Xeon® Silver 4110 CPU @ 2.10GHz, 64 GB of RAM, and an NVIDIA GeForce RTX 3090 GPU (24 GB). The experimental conditions and hyperparameter configurations were kept consistent across both the real-world and synthetic datasets. All models were trained for 200 epochs with a batch size of 2. The training process took approximately 11 hours on the synthetic dataset and 3 hours on the real-world dataset. We adopted the Adam optimizer (β1 = 0.9, β2 = 0.999) and used a cosine annealing learning rate decay strategy.

4.3 Ablation Experiment

In the ablation study, to verify the effectiveness of the proposed method, we conduct ablation experiments on the real-world dataset to evaluate the impact of the embedding position of the proposed 3D convolution fusion module.

4.3.1 Fusion3D encoder position ablation

At the encoder stage, 3D convolutional feature fusion is applied to its output features. After training the model, the RMSE, inference time, and inference time after merging the multi-branch structure are calculated. The results are presented in Table 1.

Method	*Paramenters（M）*	*RMSE(mm)*	*Time(s)*
Base	4.42	1.574	**0.008**
Base-unfold	4.51	1.489	0.010
Base-f_0-F3D	4.51	1.512	0.010
Base-$f_{0,1/16}$-F3D	4.51	1.433	0.010
Base-$f_{0,1/16,1/8}$-F3D	4.53	1.401	0.010
Base-$f_{0,1/16,1/8,1/4}$-F3D	4.54	**1.355**	0.010

Table 1: Ablation results of 3D convolutional fusion module embedded encoder.

The specific meaning of the experimental version in Table 1 is as follows:

Base: Refers to the basic model, which is the part of the network structure in Figure 2 that does not include Fusion3D.

Fusion3D.

Base-unfold: Adds the unfolding operation to the input part of the base model.

Base-f_0-F3D: Applies 3D convolution feature fusion to the output features of the last stage of the encoder.

Base-$f_{0,1/16}$-F3D: Uses 3D convolution feature fusion for the output features of the fifth and fourth stages of the encoder.

Base-$f_{0,1/16,1/8}$-F3D: Applies 3D convolution feature fusion to the output features of the fifth, fourth, and third stages of the encoder.

Base-$f_{0,1/16,1/8,1/4}$-F3D: Uses 3D convolution feature fusion for the output features of the fifth, fourth, third, and second stages of the encoder. This is the final model, 3DFNet.

Time: indicates the inference time tested when merging the multi-branch structure in the RepVGG [5] block.

The experimental results in Table 1 demonstrate that 3D convolution feature fusion effectively improves the network's performance. Additionally, the use of structural re-parameterization technology significantly reduces inference time.

4.3.2 Ablation study for SR-Layer module

The difference is that, in addition to the original upsampling branch, SR-Layer also uses an additional bilinear interpolation upsampling branch and finally uses 3D convolution to fuse the features of the two branches. The experiment calculates the RMSE after the model training, the inference time, and the inference time after the model merges the multi-branch structure. The experimental results are shown in Table 2.

The specific meaning of the experimental version in Table 2 is as follows:

Base-$f_{0,1/16}$-*SR*-F3D: 3D convolution feature fusion is used for the output features of the fifth and fourth stages of the encoder and all SR layers.

Base-$f_{0,1/16,1/8}$-*SR*-F3D: 3D convolution feature fusion is used for the output features of the fifth, fourth, and third stages of the encoder and all SR-Layers.

It can be seen that inserting a 3D convolution fusion module into the SR-Layer can also effectively improve the accuracy of model depth estimation, but the number of model parameters and inference time are significantly increased.

Method	*Paramenters（M）*	*RMSE(mm)*	*Time(s)*
Base-$f_{0,1/16}$-F3D	4.51	1.433	0.010
Base-$f_{0,1/16}$-SR-F3D	4.51	**1.405**	0.012
Base-$f_{0,1/16,1/8}$-F3D	4.53	1.401	0.010
Base-$f_{0,1/16,1/8}$-SR-F3D	4.53	**1.359**	0.013

Table 2: The results of the ablation experiment of the 3D convolutional fusion module embedded in the up-sampling module.

4.4 Model Comparison Experiment

To further demonstrate the superiority of our proposed model in the task of fringe structured light depth estimation, we train the model using an open-source real-world dataset and compare the evaluation results of published models on a test set of a real-world dataset.

As shown in Table 3, our proposed model outperforms existing models in terms of the RMSE index, while having approximately half the number of parameters. Additionally, the inference time of our model is comparable to that of existing models. Different versions of the model offer greater flexibility in memory usage and inference time, while maintaining high accuracy.

Method	*Paramenters（M）*	*RMSE(mm)*	*Time(s)*
FCN[18]	-	2.03	-
AEN[25]	-	1.85	-
Unet[28]	8.63	1.62	**0.005**
hNet[23]	8.64	1.64	**0.005**
UNet-Wavele[35]	8.64	1.67	-
hNet-Wavelet[35]	8.64	1.59	-
DHDNet[31]	14.4	1.77	-
SIDO[22]	-	1.54	0.030
Our(Base)	**4.51**	1.574	0.008
Our(Base-$f_{0,1/16,1/8}$)	4.53	1.401	0.010
Our(3DFNet)	4.53	**1.353**	0.010

Table 3: Evaluating the model on the test set of a real-world dataset.

This section also includes experiments using open-source models on the synthetic dataset. In these experiments, the depth range of the synthetic dataset is adjusted from (0, 1) to (0, 50), aligning it more closely with the depth range of the real-world dataset. The maximum PSNR value is set to 50. The experimental results are presented in Table 4.

Method	*Paramenters（M）*	*PSNR(dB)*	*MSE(mm)*	*RMSE(mm)*	*Time(s)*
Unet[28]	8.63	18.859	42.215	6.087	0.005
hNet[23]	8.64	25.334	9.079	2.845	0.005

Base-$f_{0,1/16,1/8}$	**4.53**	**28.001**	**7.538**	**2.331**	0.010

Table 4: Evaluation results of the model on the simulation dataset.

A higher PSNR indicates lower distortion in the depth image, while lower MSE and RMSE values correspond to higher depth estimation accuracy. The results above demonstrate that the method proposed in this chapter also exhibits significant advantages on the synthetic dataset.

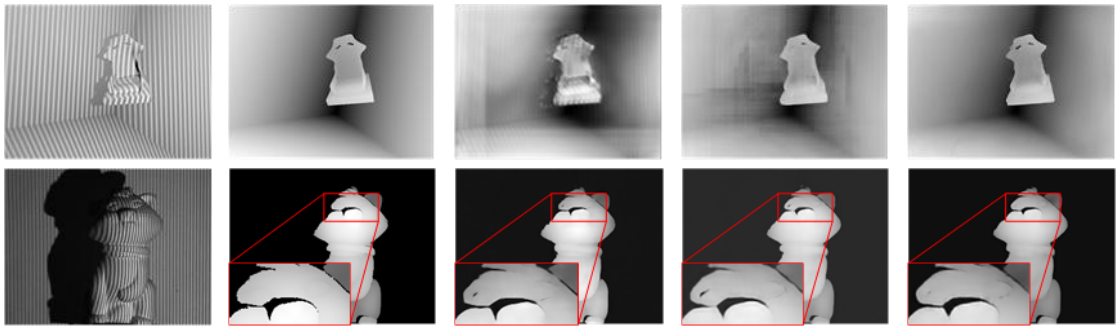
Additionally, we conducted a transfer learning experiment to evaluate how pretraining on the synthetic dataset can enhance the model's performance in real-world depth estimation tasks. Specifically, the model was first pretrained for 64 epochs on our synthetic dataset, followed by fine-tuning for 200 epochs on the real-world dataset. The final results of this experiment are presented in Table 5.

Method	*Paramenters（M）*	*Times(s)*	*RMSE(mm)*
Base-$f_{0,1/16,1/8}$	4.53	0.010	1.401
Base-$f_{0,1/16,1/8}$*+pre-trained*	4.53	0.010	**1.343**

Table 5: Performance results from pre-training on simulation datasets.

Where "+ *pre-trained*" represents the model after pre-training.

From the experimental results, it can be seen that the large simulation dataset is beneficial in improving the accuracy of fringe structured light depth estimation. By pre-training on the simulation dataset and fine-tuning on the real-world dataset, the difficult problem of collecting the real-world dataset can be effectively solved, and the accuracy of the network on the real-world dataset can be effectively improved.



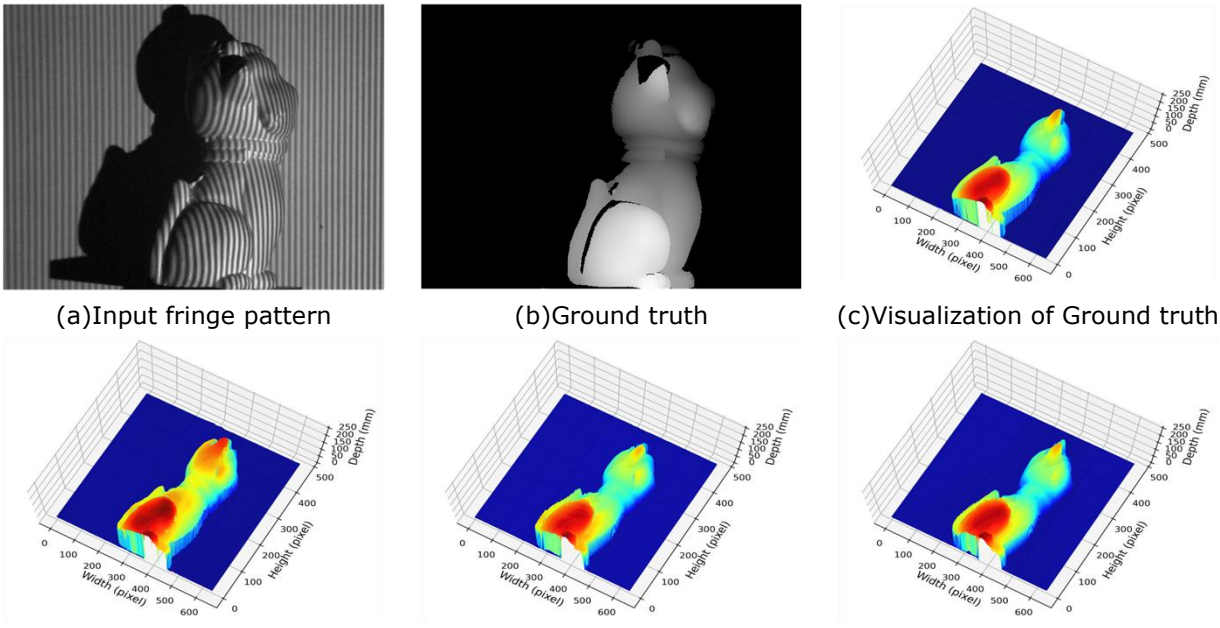
(a)Image (b)Label (c)Unet (d)hNet (e)Ours

Figure 5: Results of depth estimation of the model in the synthetic dataset and the real-world dataset. (a)The input fringe structured light image(The first row is from the synthetic dataset, and the second row is from the real-world dataset).(b)The label values of the input image. For each input image, the depth information obtained by (c) Unet [28], (d) hNet [23] , and (e) Ours (3DFnet), is illustrated above.

Finally, the qualitative analysis of the proposed algorithm shows the intuitive effect of the model proposed in this chapter and the network depth estimation of UNet [28] and hNet [23]. The results are shown in Figure 5.

It can be seen from the results that, with the increase in the amount of data, UNet and hNet can not fit well, and the predicted depth map has a large deviation in the challenging fringe coverage area, but the proposed method can restore the depth of this part of the area very well. In the real-world dataset, the three models can achieve better results, but for the areas with rich details (such as the cat's eye), the method proposed in this paper can better estimate the depth.

As shown in Figure 6, our qualitative analysis also shows the 3D visualization results of UNet[36], hNet[23], and the depth estimation of the proposed model on the real-world dataset. On the real-world 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 6) does not.



(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 our result

Figure 6: 3D visualization of model depth estimation.

5 CONCLUSIONS

In this paper, we propose a network architecture for depth estimation from structured light images. The core of the network utilizes an encoder-decoder structure, incorporating both structural re-parameterization technology and a 3D convolution feature fusion module. Structural re-parameterization enables an optimal balance between inference speed and accuracy, ensuring high performance during the inference stage. The 3D convolution feature fusion module effectively expands the receptive field. Our network takes a single-frame fringe structured light image as input and outputs the corresponding depth map. We conduct comprehensive experiments on our proposed network and other depth estimation networks across two datasets. The experimental results demonstrate that our method outperforms existing approaches in terms of parameter count, estimation accuracy, and maintains a reasonable inference speed. Additionally, in regions with rich details in the fringe structured light image, our method shows superior performance compared to other methods.

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

ACKNOWLEDGMENTS

We thank all the reviewers for their valuable comments. This research is supported by funding from the National Science Foundation of China (Nos. 61662006, 62062015).

REFERENCES

[1] Blender Foundation: Blender (3.4), 2022. https://www.blender.org/
[2] Chang, J.-R.; Chen, Y.-S.: Pyramid stereo matching network, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 5410–5418. https://doi.org/10.1109/CVPR.2018.00567
[3] Collobert, R.; Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning, Proceedings of the 25th International Conference on Machine Learning, 2008, 160–167. https://doi.org/10.1145/1390156.1390177
[4] 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
[5] 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
[6] Durrant, J.D.: Blendmol: Advanced macromolecular visualization in Blender, Bioinformatics, 35(13), 2019, 2323–2325. https://doi.org/10.1093/bioinformatics/bty968
[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. https://doi.org/10.1117/1.AP.1.2.025001
[8] 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
[9] Graves, A.; Mohamed, A.-r.; Hinton, G.: Speech recognition with deep recurrent neural networks, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947
[10] 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
[11] 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
[12] Kent, B.R.: Visualizing astronomical data with Blender, Publications of the Astronomical Society of the Pacific, 125(928), 2013, 731. https://doi.org/10.1086/671412
[13] Krizhevsky, A.; Sutskever, I.; Hinton, G.E.: ImageNet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 25, 2012.
[14] 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
[15] Li, J.; Chen, J.; Sheng, B.; Li, P.; Yang, P.; Feng, D.D.; Qi, J.: Automatic detection and classification system of domestic waste via multimodel cascaded convolutional neural network, IEEE Transactions on Industrial Informatics, 18(1), 2021, 163–173. https://doi.org/10.1109/TII.2021.3085669

[16] Liu, Y.; Yu, J.; Han, Y.: Understanding the effective receptive field in semantic image segmentation, Multimedia Tools and Applications, 77, 2018, 22159–22171. https://doi.org/10.1007/s11042-018-5704-3
[17] 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, 2021.
[18] 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
[19] 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.
[20] Marrugo, A.G.; Gao, F.; Zhang, S.: State-of-the-art active optical techniques for three-dimensional surface metrology: a review, Journal of the Optical Society of America A, 37(9), 2020, 60–77. https://doi.org/10.1364/JOSAA.398644
[21] Mehta, S.; Rastegari, M.: MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer, arXiv preprint arXiv:2110.02178, 2021.
[22] Nguyen, A.-H.; Rees, O.; Wang, Z.: Learning-based 3d imaging from single structured-light i mage, Graphical Models, 126, 2023, 101171. https://doi.org/10.1016/j.gmod.2023.101171
[23] 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
[24] Nguyen, H.; Tran, T.; Wang, Y.; Wang, Z.: Three-dimensional shape reconstruction from single-shot speckle image using deep convolutional neural networks, Optics and Lasers in Engineering, 143, 2021, 106639. https://doi.org/10.1016/j.optlaseng.2021.106639
[25] 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
[26] Puljić, A.; Zoraja, D.; Petković, T.: Simulation of structured light 3d scanning using Blende-r , 2022 International Symposium ELMAR, 2022, 215–220. https://doi.org/10.1109/ELMAR55880.2022.9899809
[27] Reitmann, S.; Neumann, L.; Jung, B.: Blainder—a Blender AI add-on for generation of semantically labeled depth-sensing data, Sensors, 21(6), 2021, 2144. https://doi.org/10.3390/s21062144
[28] 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
[29] 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
[30] Shi, J.; Zhu, X.; Wang, H.; Song, L.; Guo, Q.: Label enhanced and patch based deep learning for phase retrieval from single frame fringe pattern in fringe projection 3d measurement, Optics Express, 27(20), 2019, 28929–28943. https://doi.org/10.1364/OE.27.028929
[31] 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
[32] Wang, Z.: Review of real-time three-dimensional shape measurement techniques, Measurement, 156, 2020, 107624. https://doi.org/10.1016/j.measurement.2020.107624
[33] Yin, W.; Chen, Q.; Feng, S.; Tao, T.; Huang, L.; Trusiak, M.; Asundi, A.; Zuo, C.: Temporal phase unwrapping using deep learning, Scientific Reports, 9(1), 2019, 20175. https://doi.org/10.1038/s41598-019-56222-3
[34] Zhou, W.; Song, Y.; Qu, X.; Li, Z.; He, A.: Fourier transform profilometry based on convolution neural network, Optical Metrology and Inspection for Industrial Applications V, vol. 10819, 2018, 351–359. https://doi.org/10.1117/12.2500884

[35] 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
[36] 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
[37] Zhu, X.; Zhang, Z.; Hou, L.; Song, L.; Wang, H.: Light field structured light projection data generation with Blender, 2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications, 2022, 1249-1253. https://doi.org/10.1109/CVIDLICCEA56201.2022.9824921
[38] 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