EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

3D reconstruction from structured-light profilometry with dual-path hybrid network

Lei Wang[1,2*], Dunqiang Lu[1,2], Ruowen Qiu[1] and Jiaqing Tao[1]

*Correspondence:
wanglei2014@tju.edu.cn
[1] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, China
Full list of author information is available at the end of the article

Abstract

With the rapid development of high-speed image sensors and optical imaging technology, these have effectively promoted the improvement of non-contact 3D shape measurement. Among them, striped structured-light technology has been widely used because of its high measurement accuracy. Compared with classical methods such as Fourier transform profilometry, many deep neural networks are utilized to restore 3D shape from single-shot structured light. In actual engineering deployments, the number of learnable parameters of convolution neural network (CNN) is huge, especially for high-resolution structured-light patterns. To this end, we proposed a dual-path hybrid network based on UNet, which eliminates the deepest convolution layers to reduce the number of learnable parameters, and a swin transformer path is additionally built on the decoder to improve the global perception of this network. The experimental results show that the learnable parameters of the model are reduced by 60% compared with the UNet, and the measurement accuracy is not degraded at the same time. The proposed dual-path hybrid network provides an effective solution for structured-light 3D reconstruction and its practice in engineering.

Keywords: Structured-light, 3D reconstruction, Profilometry, Deep neural network, Dual-path decoder, Swin transformer

1 Introduction

As a high-accuracy non-contact 3D reconstruction method, the striped structured-light profilometry has been widely employed in geometric measurement, relic restoration, reverse engineering, etc. [1, 2]. Structured light measurement has also been applied in automatic driving in recent years, while the real-time perception of vehicles and obstacles is the research hotspot of IOV (Internet of vehicle) and an important prerequisite for automatic and unmanned driving [3]. For road environment perception in automobiles, LIDAR and camera are the most widely used sensing devices. LIDAR has a long detection distance, but the resolution of 3D imaging is low, and the high cost restricts its application, while a single camera without auxiliary lighting cannot calculate the 3D profile of the road environment [4, 5]. With the help of invisible light sources, a single camera can also carry out road environment reconstruction based on the structured-light profilometry.

It projects a structured-light pattern onto the surface to be measured, and then the pattern is modulated by the object surface, and the camera collects the deformed structured-light pattern and restores the height map by demodulating the phase information of the deformed structured-light pattern [6, 7]. The Fourier transform profilometry is a classical structured-light phase demodulation algorithm, it converts the structured-light pattern from the spatial domain to the frequency domain and removes the high-frequency component and 0 Hz component in the frequency domain with an appropriate filter, and the rest of the fundamental frequency components are restored to the spatial domain by IFFT (inverse fast Fourier transform) [8]. Then, the phase information is demodulated according to the complex form of the restored structured-light pattern [9]. However, the demodulated phase is wrapped in the range of $(-\pi, \pi)$, and auxiliary methods such as spatial information are needed, so that the phase can be unwrapped to calculate the 3D shape of the measured object [10].

In engineering applications, Fourier transform profilometry can reconstruct 3D profile with a single-shot structured-light pattern, so it has the advantage of high measurement speed, which is convenient for online measurement and dynamic measurement. However, the filter parameter setting and phase unwrapping algorithm of the Fourier transform method are difficult, which limit its accuracy and robustness.

In this work, a novel dual-path hybrid model is proposed for structured-light profilometry. Based on UNet, this approach deletes the deepest convolution layers in the neural networks to reduce the number of learnable parameters, and a swin transformer path is added at the decoding end to improve the global perception ability of the model. Experimental results show that this method can reduce the size of the model and reconstruct 3D profile with high accuracy [11].

2 Related work

In the theoretical research of structured-light 3D profilometry, for non-contact, high-precision measurement of pavement texture, Wang et al. [12] developed an innovative surface structured-light projection (SSLP) based on optical fiber interference, the wrapped phase can be demodulated accurately from the wavelet ridge by two-dimensional continuous wavelet transform, and then the measured pavement texture elevation can be calculated according to the phase-height mapping relationship. A combined approach to improve 3D object shape recovery based on Fourier orthogonal structured-light pattern projection together with Hilbert transform is proposed in [13], which can suppress the background intensity of the deformed fringe pattern, and experimental results verified better performance in reconstruction of complex objects. In [14], they construct a 2D continuous complex wavelet employing a 2D real Mexican hat wavelet function, combined with the single-orthant analytical 2D Hilbert transform, and the experiments demonstrate that it provides high phase accuracy in the single-shot fringe pattern profilometry.

There are also many research improvements on lighting hardware and structured-light pattern design nowadays. In [15], for the application of structured-light profilometry on a microscopic scale, they present a Gates' interferometer configuration with an LED source to project a structured-light pattern without speckle noise and a very long field depth, and the system can obtain excellent sinusoidal structured

light on the surface of microscopic objects. Linear structured-light patterns cannot uniquely represent the lateral displacement caused by objects with surface discontinuities, Mandapalli et al. [16] propose using a radially symmetric circular structured-light pattern as the structured-light pattern for accurate unambiguous surface profiling of sudden height-discontinuous objects, and experimental results prove that the proposed method can be applied for the reconstruction of objects with 4 times higher dynamic range and even at much lower structured-light frequencies.

Nevertheless, the traditional structured-light 3D profilometry is troublesome for effective 3D reconstruction in engineering applications. At present, deep neural networks are widely used in image processing [17–19], so it has become a new research hotspot to apply depth learning to the analysis of structured-light patterns for fast 3D profile measurement. Many scholars have conducted preliminary research on structured-light 3D profilometry with depth learning. Plenty of early studies conduct neural networks for one of the steps in structured-light profilometry, such as structured-light pattern denoising, phase extraction and phase unwrapping. In [20], two low-modulation patterns with different phase shifts are transformed into a set of three phase-shifted high-modulation fringes by using FMENet. Yu et al. [21] propose a novel phase retrieval technique based on CNN, which uses an end-to-end deep convolution neural network to transform a single or two patterns into the phase retrieval required patterns, and numerically and experimentally verified its applicability for dynamic 3D measurement. In [22], they employ UNet to transform a color structured-light pattern into multiple triple-frequency phase-shifted grayscale patterns, from which the 3D shape can be accurately reconstructed. Now, there are many works on end-to-end height map directly. In [23], a network with 10 convolutional layers is built for full-field height extraction from structured-light pattern. Qiao et al. [24] utilize depth-wise separable convolution to build a deep neural network, which can reduce the number of learnable parameters of the model, and the accuracy of 3D reconstruction does not decrease. In the structural design of CNN, there are many research works based on UNet. Nguyen et al. [25] compare different types of end-to-end networks, and the experimental results demonstrate the high accuracy of UNet reconstruction results. Nguyen et al. [26] use an end-to-end neural network to reconstruct the 3D profile by transforming a single speckle-pattern image into its corresponding 3D point cloud.

In addition to the end-to-end full convolution network (FCN), there are plenty of studies utilizing the structure of multi-path neural networks. Qiao et al. [24] present a multi-path CNN to predict the high-resolution, crosstalk-free absolute phase directly from one single color fringe image, which allows for more accurate phase retrieval and more robust phase unwrapping. For different types of multi-path CNN models, Cywinska et al. [27] evaluate the effects of the number of paths and the number of filters on the RMSE (root-mean-square error) of the reconstruction result, and also time consumption; then, they give recommended parameters. Nguyen et al. [28] transform multiple (typically two) grayscale images consisting of fringe and/or speckle patterns into a 3D depth map using a multi-path neural network and fuse multiple feature maps to obtain multiple outputs with an accuracy-enhanced final output. In [29], a novel dual-dense block structure is designed and embedded into a multi-path structure to fully utilize the

local layers and fuse multiple discrete sinusoidal signals, with which highly reconstruction results can be obtained even when training with a smaller data sample.

It should be noted that although there are many studies on structured-light 3D measurement based on deep learning, limited by the number of layers of deep neural networks, there are few lightweight models for high-resolution 3D reconstruction, and the weak long-distance interaction ability of convolution also affects the reconstruction accuracy of networks [30, 31].

3 Methods

For the principle of different structured-light profilometries, Fourier transform profilometry has the lowest computational complexity, and the deformed structured-light pattern $I(u, v)$ captured by camera can be written as:

$$I(u,v) = a(u,v) + b(u,v)\cos[\varphi(u,v) + 2\pi f_0 u] \tag{1}$$

where $a(u, v)$ is the background light intensity of pixels (u, v), $b(u, v)$ is the amplitude of structured-light pattern, f_0 is the fundamental frequency of the striped structured-light, and $\varphi(u, v)$ is the phase amplitude modulated by the surface height $h(u, v)$. Then, $h(u, v)$ is expressed as:

$$h(u,v) = \frac{l_0\varphi(u,v)}{2\pi f_0 d} \tag{2}$$

where d represents the central distance between the camera and projector, and l_0 represents the distance between reference plane and the camera, and both of them are the geometric parameters of the structured-light device.

Convert the trigonometric function in Eq. 1 into an exponential form, let $c(u,v) = \frac{1}{2}b(u,v)\exp(i\varphi(u,v))$, the phase amplitude $\varphi(u, v)$ modulated by the measured surface can be expressed as:

$$\varphi(u,v) = \frac{Im[c(u,v)\exp(i2\pi f_0 u)]}{Re[c(u,v)\exp(i2\pi f_0 u)]} \tag{3}$$

The phase amplitude $\varphi(u, v)$ calculated by Eq. 3 is wrapped in the range of $(-\pi, \pi)$; after phase unwrapping, the final height map $h(u, v)$ can be obtained by Eq. 2 [32].

3.1 Global feature extraction of structured-light pattern

At present, for structured-light measurement algorithms based on deep neural network, most of them are encoder–decoder frameworks. Feature maps are extracted from the input structured-light pattern by a pre-trained network and then put them into the decoder to generate height information.

In the encoder, for feature map extraction of structured-light pattern, it is necessary to collect global features, especially when there are discontinuous sections of the measured surface. The current approaches are: (1) By reducing the resolution (reducing the scale) of the convolution layer feature map, such as down-sampling operations (e.g., pooling layer), the network can get the feature information between long-distance positions of the original pattern. However, the output of the convolution layer represents the feature information at different spatial positions, and the pattern is segmented into

grids to obtain the local features of each part. In the process of encoding and decoding, the scaling of pattern size leads to the loss of information, resulting in the reduction of the accuracy of the 3D reconstruction, or relying on deeper convolution and pooling operations [33]. (2) Another method is dilated convolution, compared with the problem that the pooling layer increases the receptive field but losses information, and dilated convolution network can avoid the down-sampling operation [34]. By adding a dilation rate, dilated convolution inserts blanks between the elements of the convolution kernel, which expands the kernel for a larger receptive field. However, the sampling process of dilated convolution is sparse, while multiple dilated convolutions are superimposed in the network, and some lost pixels will lose the continuity of information and the correlation between the feature maps, for object edge and small scale object, which will result in the decrease of 3D reconstruction accuracy [35].

Nowadays, most of existing studies are based on deep convolution layers to extract global feature maps of the structured-light pattern, which leads a large number of learnable parameters of the network, long training time and difficult deployment. Therefore, for efficient and accurate 3D reconstruction, a key step is to get more global information based on the network with limited neural layers [36, 37].

For global feature maps extraction, self-attention makes great improvement in acquiring large-scale interactivity, which main operation is to obtain the weighted average of the calculated values of hidden cells. More than that, self-attention mechanism can get a wide range of interactive without increasing parameters, which helps to reduce the number of learnable parameters of the network model. This is significant for large-scale modeling of high-resolution structured-light profilometry [38, 39].

At present, transformer uses self-attention to acquire long-range interactive information. Compared with CNN, the transformer requires fewer computing resources, has achieved excellent performance in NLP, image classification, etc. [40, 41], and has become a study hotspot in deep learning. The underlying structure of the transformer is similar to ResNet, which divides the image into multiple patches of a specified size, and this leads to two disadvantages: First, the boundary pixels cannot use the adjacent pixels outside the patch for image restoration; second, the restored image may be mixed with boundary artifacts around each patch [42].

As an improved visual transformer, swin transformer utilizes a novel general architecture based on shifted-window and hierarchical expression. Compared with the previous vision transformer, swin transformer introduces the idea of locality and uses the shifted window to calculate the self-attention of the non-coincident patches, which also greatly reduces the computing consumption [43, 44].

3.2 Dual-path hybrid submodule

Convolution has good local perception ability, but it lacks the interaction of long-range information, which will lose the global feature of the structured-light pattern. If the network only relies on deeper convolution layers and pooling layers to expand the receptive field, it will lead to a huge number of learnable parameters and over-fitting of the network. A pure transformer or swin transformer network has an obvious advantage in the global perception of the pattern, but the pattern detail information is lost in the division of patches [45]. In [46], a hybrid network structure is proposed to take advantage

of convolutional operations and self-attention mechanisms for enhanced representation learning, which can significantly improve the representation ability of the base network under comparable parameter complexity. Inspired by this, we present a dual-path hybrid submodule for feature learning, in which there are two parallel subpaths, the local and global features are represented by convolution path and swin transformer path, respectively, and each convolution block has its corresponding parallel swin transformer block for feature interaction [47]. The diagram of the dual-path hybrid submodule is shown in Fig. 1.

In the convolution path of the dual-path hybrid submodule, the feature map f_i output by the previous submodule is directly transmitted to the convolution path for local feature extraction, and this feature is also serialized by a FC (feature coupling) Down block and indirectly sent to the swin transformer path for global feature extraction. The output global feature p_s of the swin transformer is converted into 3D form f_u ($H_jW_jC_j$) by a FC Up block, and it is coupled with the output feature f_c from the convolution layer by the Average layer. There are a UpSampling2D layer and a Dense layer in the behind of the Average layer, and the purpose is to keep the feature dimension consistent with the residual information from the encoder. After the feature information and residual information are concatenated, they are used as the input f_j for the next submodule [48, 49].

In the swin transformer path, the tensor p_i output by the previous submodule and the 2D feature map passed from the convolutional layer are also coupled by the Average layer and then passed to the current swin transformer block for global feature extraction. The tensor p_s gets from the swin transformer has two branches: One is coupled to the convolution path for providing global feature information, and the other is upsampled by patch expanding layer and passed to the next submodule for further global feature representation.

The FC Down block is composed of a patch extracting layer, a patch embedding layer and a LayerNormalization layer, and the 3D feature map f_i is serialized by the patch extracting layer into 2D patches by the patch extracting layer. These patches are tokenized by the patch embedding layer and maintain a similar dimension to the previous p_i; after the LayerNormalization layer, the disappearance of gradient can be avoided. In the FC Up block, after a patch expanding 2D layer, the serialized global feature p_s is reshaped into 3D form; then, its dimension is supplemented by the 11 convolution layer and then outputs through the BatchNormalization layer.

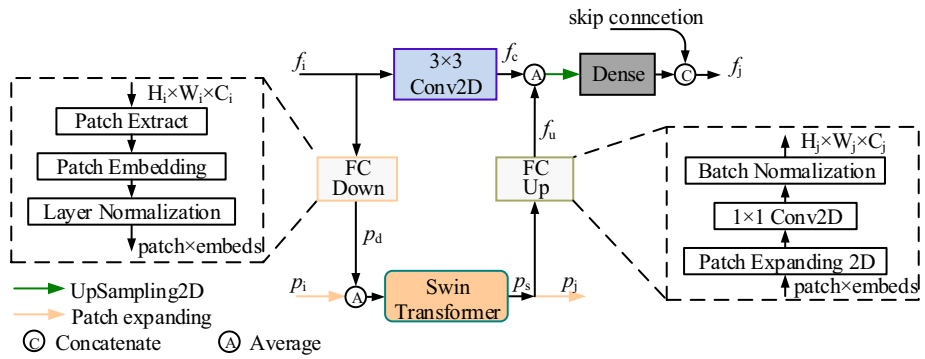

Fig. 1 The diagram of the dual-path hybrid submodule

With this dual-path hybrid submodule, for feature maps with different scales, the convolutional path and the swin transformer path can extract local and global features, respectively, and those two different features are strongly fused by coupling blocks. Through the hybrid submodule, the number of layers of the neural network can be effectively reduced, and a high-precision 3D reconstruction can be obtained.

3.3 The proposed dual-path hybrid decoder network

Based on the dual-path hybrid submodule mentioned above, we proposed a novel dual-path decoder network for single-shot structured-light profilometry, which is improved from the classic UNet [50], and the final network architecture is shown in Fig. 2.

There are three convolution blocks in the encoder, which consist of two 33 convolution layers, two BatchNormalization layers and a MaxPooling layer. Between the convolution blocks, 2 down-sampling is performed by the MaxPooling layer. Compared with 4 downsampling convolution blocks and a bottom convolution block in UNet, the proposed network eliminates the deepest convolution blocks in order to reduce the overall size of the model, and the global feature information is extracted and represented by the hybrid submodules in the decoder. Meanwhile, each convolution block also outputs residual information for skipping to the decoder, which can avoid the gradient disappearance in the back-propagation process [51].

The decoder is composed of 4 dual-path hybrid submodules in series, which mainly represent the local and global features of the structured-light pattern, and scale the feature maps in the two paths by UpSampling layer and Patch expanding layer, respectively. It should be noted that in the decoder, each convolution block consists of one 33 convolution layer and one BatchNormalization layer, while each swin transformer block is composed of two swing transformer layers.

The output layer in the model is a 11 convolution, and the final 3D height map is output in the form of linear regression.

4 Experiments and analysis

In order to verify the effectiveness of the proposed network in structured-light profilometry, we compare our method with the existed methods, including classical UNet and different ablation models. Sufficient experiments have proved the feasibility and lightweight of this method in 3D reconstruction.

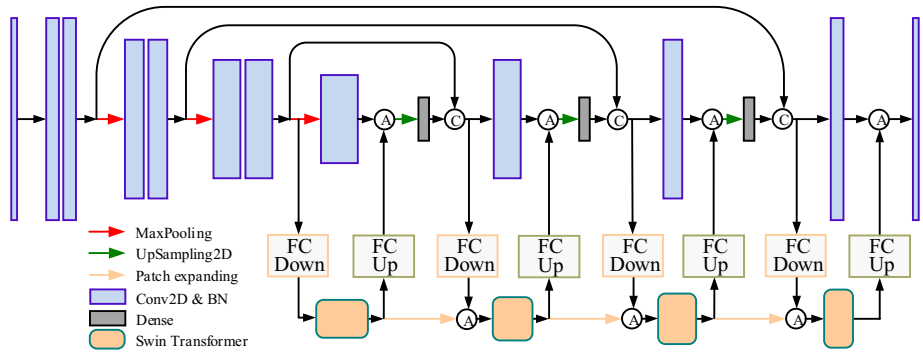

Fig. 2 Structure diagram of the proposed network

The experimental hardware platform is a server that consists of an NVIDIA Tesla P100 GPU, 64GB ram and an Intel Xeon 4110 CPU. The software tools that we select are Python 3.8 and deep learning framework TensorFlow.

4.1 Visual assessment and accuracy evaluation

For the selection of dataset, we use the actual dataset that mentioned in [25], which is collected from the structured-light patterns of real gypsum sculptures, and the ground truths are measured by phase-shift method. In order to expand the number of samples, the dataset is expanded by rotating and translating these gypsum sculptures randomly.

The total number of structured-light patterns in training set is 500, and the number of validation set is 100. For the training of UNet and the proposed network, let batch size = 2, the initial learning rate is set to 0.0001. After 200 epochs training, we randomly select structured-light patterns in the test set for 3D reconstruction prediction. The final visual results are shown in Fig. 3.

For the visual performance of the comparison experiments in Fig. 3, the 1st column is ground truths of sculptures, the 3D reconstruction results of UNet are in the 2nd column, and the 3rd column is the results of the proposed network. In our subjective analysis, compared with Fig. 3b, most of the folds of the clothing in Fig. 3c can be exposed. This shows that it is useful to add global features to the convolution path, so the proposed network can obtain richer height map details and fully display the overall effect, which can effectively improve the accuracy of 3D reconstruction.

In order to quantitatively compare the 3D reconstruction errors of different networks, we choose to compare the measurement errors of Fig. 3b and c. We take the ground truth of the height map measured by the phase-shift method as the reference and carry out the least square fitting on the reconstruction results of the two networks, so as to scale the reconstruction height map in the range of (0, 255). The final absolute errors of the different networks are shown below:

Figure 4 shows the error comparison of different networks, and the height map of UNet is partially offset from the ground truth after surface fitting in Fig. 4a. The average

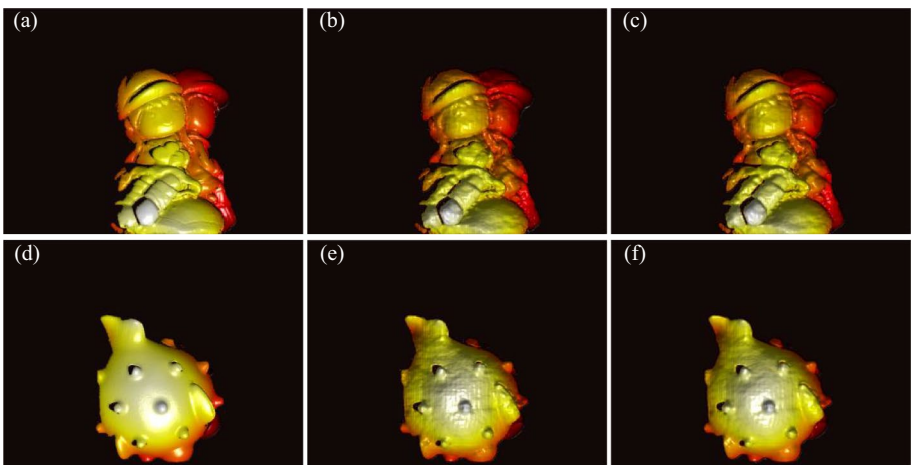

Fig. 3 3D reconstruction of different sculptures: **a**, **d** ground truth; **b**, **e** results of UNet; **c**, **f** results of the proposed network

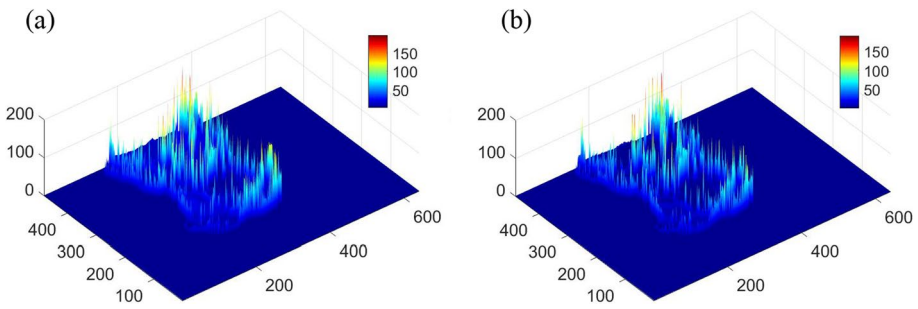

Fig. 4 Measurement error of different networks: **a** UNet; **b** the proposed network

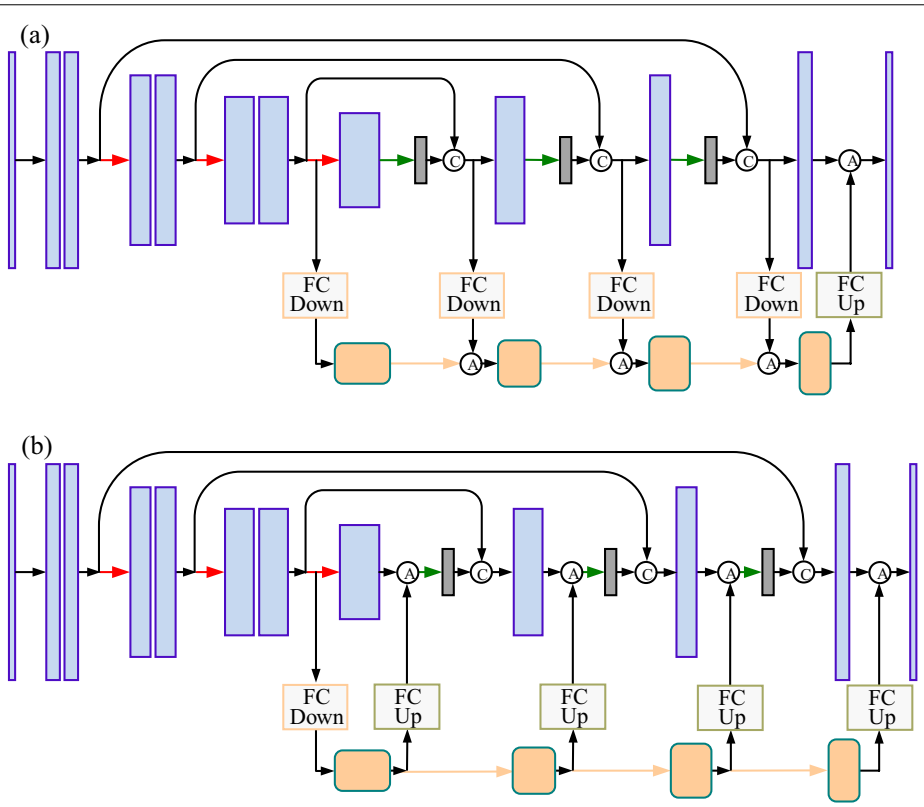

Fig. 5 Structural comparison of simplified dual-path hybrid networks: **a** Hybrid-A; **b** Hybrid-B

error of UNet is 8.34 RMSE/pixel, and the proposed network is 7.79 RMSE/pixel; for the maximum error, the proposed network is reduced by 8% compared with UNet.

4.2 Ablation experiment

More than that, in order to verify the importance of mutual fusion of local features and global features in the hybrid network, we design two simplified unidirectional feature fusion networks and compared with the proposed network. The diagrams of those two simplified dual-path hybrid networks are shown in Fig. 5.

In the unidirectional feature fusion network A (Hybrid-A), compared with the previous proposed network, we delete the FC Up block in some hybrid submodules. Only in

Table 1 The performance comparison of networks with different structures

	Learnable parameters (MB)	Normalized MAE
UNet	359	0.0092
Proposed network	144	0.0087
Hybrid-A	141	0.0088
Hybrid-B	99	0.0102

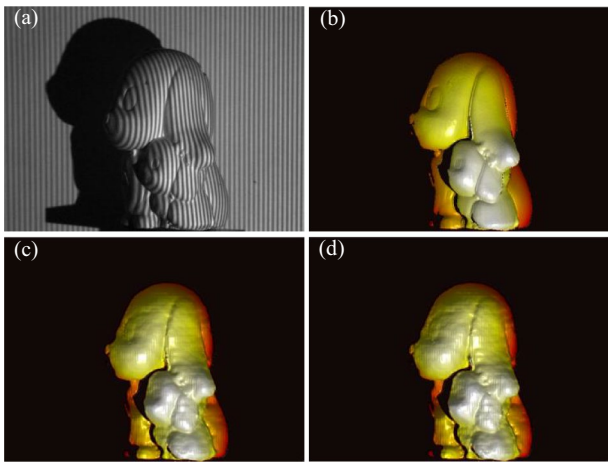

Fig. 6 The performance comparison of simplified dual-path hybrid networks: **a** the structured-light pattern; **b** ground truth; **c** result of Hybrid-A; **d** result of Hybrid-B

the last hybrid submodule, the global feature information of the swin transformer path is fused back to the convolution path by the FC Up block, while other hybrid submodules only fuse the local feature information from the convolution path through the FC Down block to the swin transformer path.

For the other unidirectional feature fusion network B (Hybrid-B), which is similar to Hybrid-A on the contrary, in the first hybrid submodule, the local feature information of the convolution path is fused to the swin transformer path by the FC Down block, and the subsequent hybrid submodules only transfer the global feature information back to the convolution path from the swin transformer path, but no local feature information is transferred to the swin transformer path.

Those two simplified networks are trained on the same dataset, and the quantitative comparison results with the proposed network and UNet are shown in Table 1.

By the normalized MAE of reconstruction results of different networks, we can find out that the performance of the proposed network and Hybrid-A is better than that of UNet, while Hybrid-A is the worst. The visual performance comparison of those two simplified dual-path hybrid networks in Fig. 6, Hybrid-B has more distortion details than Hybrid-A.

5 Results and discussion

Compared with UNet, the learnable parameters of the proposed network are 60% less than UNet, and the parameters of the two simplified networks are also much less than UNet, which helps to reduce the computing consumption and the difficulty of hardware deployment.

In the structure comparing of Hybrid-A and Hybrid-B, we find out that if the deep feature map is extracted only once from the convolution path for global feature representation, even if the global features are fed back to the convolution path for many times, the back-propagation of the network still has the problem of gradient disappearance, which leads to the weak generalization capability of the hybrid network, and also the decline of the accuracy of 3D reconstruction. In Hybrid-A, the swin transformer path can extract local feature information from the convolution path repeatedly and fuses them with the global features from the previous swin transformer layer, just like the residual skip connection of ResNet, which can avoid the gradient disappearance of the model and help to improve the 3D reconstruction accuracy [52, 53].

From the comparative experiments of the above two dual-path hybrid networks, for the fusion of global and local feature information in deep neural network, we can conclude that, compared with the global features obtained unidirectionally from the swing transformer blocks, the decoder can obtain local feature information of different scales from convolution path at different layers, which can better improve the prediction performance. This also proves the positive role of multi-attention mechanism such as swin transformer in structured-light 3D reconstruction.

6 Conclusion

In order to design a lightweight structured-light 3D reconstruction network, we proposed a dual-path hybrid network based on research of multi-attention mechanism. Compared with the classical UNet, we eliminate the deepest convolution block to reduce the total learning parameters of the network. Meanwhile, to improve the global perception ability of the network, a swin transformer path is added to the decoder for global feature representation, and the local features of the convolution path are strongly fused by the bidirectional fusion submodule. The experimental result demonstrates that the learnable parameters of the proposed network are 60% less than that of UNet. For the fusion direction between the local features of the convolution path and the global features of the swin transformer path, its influence on the generalization ability of the model is verified by two simplified hybrid networks in the ablation experiment. Through these experiments, this dual-path hybrid network framework provides a new idea for structured-light 3D reconstruction and engineering applications in automatic driving.

Abbreviations

CNN: Convolution neural network; IFFT: Inverse fast Fourier transform; SSLP: Surface structured-light projection; FCN: Full convolution network; RMSE: Root-mean-square error; FC: Feature coupling; MAE: Mean absolute error; MSE: Mean squared error; SSIM: Structural similarity; GSSIM: Gradient-based structural similarity.

Acknowledgements

Not applicable.

Authors' contributions

All authors read and approved the final manuscript.

Funding

This study was supported by Tianjin Educational Commission Scientific Research Program (2020KJ004) and in part by the Doctoral Foundation of Tianjin Normal University (52XB1906).

Availability of data and materials

All the datasets used for training the model of this paper are from Internet.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin, China. [2]State Key Laboratory of Precision Measuring Technology and Instruments, Tianjin University, Tianjin, China.

Received: 17 January 2022 Accepted: 6 February 2022
Published online: 21 February 2022

References

1. P. Zhang, Z. Kai, Z. Li, X. Jin, B. Li, C. Wang, Y. Shi, High dynamic range 3d measurement based on structured light: a review **1**(2), 2021004–2021009. https://doi.org/10.51393/j.jamst.2021004
2. S. Van der Jeught, J.J.J. Dirckx, Real-time structured light profilometry: a review **87**, 18–31. https://doi.org/10.1016/j.optlaseng.2016.01.011
3. X. Liu, X. Zhang, Rate and energy efficiency improvements for 5g-based iot with simultaneous transfer. IEEE Internet Things J. **6**(4), 5971–5980 (2018)
4. X. Liu, X.B. Zhai, W. Lu, C. Wu, QoS-guarantee resource allocation for multibeam satellite industrial internet of things with NOMA. IEEE Trans. Ind. Inf. **17**(3), 2052–2061 (2019)
5. F. Li, K.-Y. Lam, X. Liu, J. Wang, K. Zhao, L. Wang, Joint pricing and power allocation for multibeam satellite systems with dynamic game model. IEEE Trans. Veh. Technol. **67**(3), 2398–2408 (2017)
6. Y. Liu, Y. Fu, Y. Zhuan, K. Zhong, B. Guan, High dynamic range real-time 3d measurement based on Fourier transform profilometry **138**, 106833. https://doi.org/10.1016/j.optlastec.2020.106833
7. H. Nguyen, J. Liang, Y. Wang, Z. Wang, Accuracy assessment of fringe projection profilometry and digital image correlation techniques for three-dimensional shape measurements **3**(1), 014004. https://doi.org/10.1088/2515-7647/abcbe4
8. X. Liu, X. Zhang, M. Jia, L. Fan, W. Lu, X. Zhai, 5g-based green broadband communication system design with simultaneous wireless information and power transfer. Phys. Commun. **28**, 130–137 (2018)
9. Z. Wu, W. Guo, L. Lu, Q. Zhang, Generalized phase unwrapping method that avoids jump errors for fringe projection profilometry **29**(17), 27181–27192. https://doi.org/10.1364/OE.436116
10. P. Lafiosca, I.-S. Fan, N.P. Avdelidis, Automated aircraft dent inspection via a modified fourier transform profilometry algorithm **22**(2), 433. https://doi.org/10.3390/s22020433
11. R. Liu, W. Cai, G. Li, X. Ning, Y. Jiang, Hybrid dilated convolution guided feature filtering and enhancement strategy for hyperspectral image classification
12. H. Wang, J. Ma, H. Yang, F. Sun, Y. Wei, L. Wang, Development of three-dimensional pavement texture measurement technique using surface structured light projection **185**, 110003. https://doi.org/10.1016/j.measurement.2021.110003
13. O.I. Rosenberg, D. Abookasis, Application of Hilbert analysis in orthogonal Fourier fringe-projection to improve object shape reconstruction. https://doi.org/10.1134/S0030400X21050131
14. M. Han, W. Chen, Two-dimensional complex wavelet with directional selectivity used in fringe projection profilometry **46**(15), 3653–3656. https://doi.org/10.1364/OL.420460
15. J. Ruben Sanchez, A. Martinez-Garcia, J. Antonio Rayas, M. Leon-Rodriguez, LED source interferometer for microscopic fringe projection profilometry using a gates' interferometer configuration **149**, 106822. https://doi.org/10.1016/j.optlaseng.2021.106822
16. J.K. Mandapalli, V. Ravi, S.S. Gorthi, S. Gorthi, R.K. Gorthi, Single-shot circular fringe projection for the profiling of objects having surface discontinuities **38**(10), 1471–1482. https://doi.org/10.1364/JOSAA.430981
17. J. Huang, S.-S. Huang, H. Song, S.-M. Hu, Di-fusion: online implicit 3d reconstruction with deep priors, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8932–8941 (2021)
18. R. Chabra, J.E. Lenssen, E. Ilg, T. Schmidt, J. Straub, S. Lovegrove, R. Newcombe, Deep local shapes: learning local SDF priors for detailed 3d reconstruction, in *European Conference on Computer Vision*, pp. 608–625 (Springer, 2020)
19. P. Dou, S.K. Shah, I.A. Kakadiaris, End-to-end 3d face reconstruction with deep neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5908–5917 (2017)
20. H. Yu, D. Zheng, J. Fu, Y. Zhang, C. Zuo, J. Han, Deep learning-based fringe modulation-enhancing method for accurate fringe projection profilometry **28**(15), 21692–21703. https://doi.org/10.1364/OE.398492
21. H. Yu, X. Chen, Z. Zhang, C. Zuo, Y. Zhang, D. Zheng, J. Han, Dynamic 3-d measurement based on fringe-to-fringe transformation using deep learning **28**(7), 9405–9418. https://doi.org/10.1364/OE.387215
22. H. Nguyen, Z. Wang, Accurate 3d shape reconstruction from single structured-light image via fringe-to-fringe network **8**(11), 459. https://doi.org/10.3390/photonics8110459
23. S. Van der Jeught, J.J.J. Dirckx, Deep neural networks for single shot structured light profilometry **27**(12), 17091–17101. https://doi.org/10.1364/OE.27.017091
24. G. Qiao, Y. Huang, Y. Song, H. Yue, Y. Liu, A single-shot phase retrieval method for phase measuring deflectometry based on deep learning **476**, 126303. https://doi.org/10.1016/j.optcom.2020.126303

25. H. Nguyen, Y. Wang, Z. Wang, Single-shot 3d shape reconstruction using structured light and deep convolutional neural networks **20**(13), 3718. https://doi.org/10.3390/s20133718
26. H. Nguyen, T. Tran, Y. Wang, Z. Wang, Three-dimensional shape reconstruction from single-shot speckle image using deep convolutional neural networks **143**, 106639. https://doi.org/10.1016/j.optlaseng.2021.106639
27. M. Cywinska, F. Brzeski, W. Krajnik, K. Patorski, C. Zuo, M. Trusiak, DeepDensity: convolutional neural network based estimation of local fringe pattern density **145**, 106675. https://doi.org/10.1016/j.optlaseng.2021.106675
28. H. Nguyen, K.L. Ly, T. Nguyen, Y. Wang, Z. Wang, MIMONet: Structured light 3d shape reconstruction by a multi-input multi-output network **60**(17), 5134–5144. https://doi.org/10.1364/AO.426189
29. P.Yao, S. Gai, F. Da, Super-resolution technique for dense 3d reconstruction in fringe projection profilometry **46**(18), 4442–4445. https://doi.org/10.1364/OL.431676
30. W. Luo, Y. Li, R. Urtasun, R. Zemel, Understanding the effective receptive field in deep convolutional neural networks, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 4905–4913 (2016)
31. A. Kazi, S. Shekarforoush, S.A. Krishna, H. Burwinkel, G. Vivar, K. Kortüm, S.-A. Ahmadi, S. Albarqouni, N. Navab, Inceptiongcn: receptive field aware graph convolutional network for disease prediction, in *International Conference on Information Processing in Medical Imaging*, pp. 73–85 (Springer, 2019)
32. M. Agnès, C. Pablo, P. Vincent, P. Philippe, Experimental and theoretical inspection of the phase-to-height relation in Fourier transform profilometry
33. M. Samy, K. Amer, K. Eissa, M. Shaker, M. ElHelw, Nu-net: deep residual wide field of view convolutional neural network for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 267–271 (2018)
34. L. Zhou, C. Zhang, M. Wu, D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 182–186 (2018)
35. Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, T.S. Huang, Revisiting dilated convolution: a simple approach for weakly-and semi-supervised semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7268–7277 (2018)
36. B. Jia, W. Feng, M. Zhu, Obstacle detection in single images with deep neural networks **10**(6), 1033–1040. https://doi.org/10.1007/s11760-015-0855-4
37. X. Liu, X. Zhang, Noma-based resource allocation for cluster-based cognitive industrial internet of things. IEEE Trans. Ind. Inf. **16**(8), 5379–5388 (2019)
38. A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, J. Shlens, Scaling local self-attention for parameter efficient visual backbones, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12894–12904 (2021)
39. H. Zhao, J. Jia, V. Koltun, Exploring self-attention for image recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076–10085 (2020)
40. H. Wu, G.Q. Shen, X. Lin, M. Li, C.Z. Li, A transformer-based deep learning model for recognizing communication-oriented entities from patents of ICT in construction **125**, 103608. https://doi.org/10.1016/j.autcon.2021.103608
41. Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, C. Xu, Transformers in computational visual media: a survey **8**(1), 33–62. https://doi.org/10.1007/s41095-021-0247-3
42. E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, A. Dai, Learning the graphical structure of electronic health records with graph convolutional transformer **34**(1), 606–613. https://doi.org/10.1609/aaai.v34i01.5400
43. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows. version: 2. 2103.14030
44. Q. Zhu, Y. Zhong, Y. Liu, L. Zhang, D. Li, A deep-local-global feature fusion framework for high spatial resolution imagery scene classification **10**(4), 568. https://doi.org/10.3390/rs10040568
45. R. Ranftl, A. Bochkovskiy, V. Koltun, Vision transformers for dense prediction, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188 (2021)
46. Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, Q. Ye, Conformer: local features coupling global representations for visual recognition
47. J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, R. Timofte, SwinIR: image restoration using swin transformer, pp. 1833–1844
48. S. Chaib, H. Liu, Y. Gu, H. Yao, Deep feature fusion for VHR remote sensing scene classification. https://doi.org/10.1109/TGRS.2017.2700322
49. Z. Zhang , X. Zhang, C. Peng, X. Xue, J. Sun, Exfuse: enhancing feature fusion for semantic segmentation, in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–284 (2018)
50. O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241 (Springer, 2015)
51. M. Nie, Z. Lei, Hybrid CTC/attention architecture with self-attention and convolution hybrid encoder for speech recognition **1549**(5), 052034. https://doi.org/10.1088/1742-6596/1549/5/052034
52. T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, R. Girshick, Early convolutions help transformers see better, in *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/2106.14881
53. L. Jia, M. Gong, Q. Kai, P. Zhang, A deep convolutional coupling network for change detection based on heterogeneous optical and radar images **PP**(99), 1–15. https://doi.org/10.1109/tnnls.2016.2636227

Publisher's Note