SCIENCE CHINA
Information Sciences

• LETTER •

January 2025, Vol. 68, Iss. 1, 119102:1–119102:2
https://doi.org/10.1007/s11432-024-4212-2

Multi-receptive field interaction network for shape from polarization

Yini PENG[1†], Rui LIU[1†], Zhiyuan ZHANG[1], Zhongyuan WANG[2], Jiayi MA[1] & Xin TIAN[1,3*]

[1]*Electronic Information School, Wuhan University, Wuhan 430072, China;*
[2]*Computer Science School, Wuhan University, Wuhan 430072, China;*
[3]*Wuhan Institute of Quantum Technology, Wuhan 430206, China*

Received 9 May 2024/Revised 5 August 2024/Accepted 27 October 2024/Published online 19 November 2024

Citation Peng Y N, Liu R, Zhang Z Y, et al. Multi-receptive field interaction network for shape from polarization. Sci China Inf Sci, 2025, 68(1): 119102, https://doi.org/10.1007/s11432-024-4212-2

Shape from polarization (SfP) method can use the polarization information in reflected light to estimate the surface normal of the target, which can further reconstruct the shape of the object. With a simple image capture process, it can use low-cost equipment to meet a high precision imaging requirement, which can be used in remote scenes and other applications.

However, there are two main difficulties in SfP, i.e., the π-ambiguity in the solution of azimuth angle and the uncertainty of reflection type. These problems could lead to low precision and convex/concave ambiguity in the reconstruction results. To solve these problems, previous researchers have explored various cues and techniques, which can be mainly divided into physics-based methods and data-driven methods. The physics-based studies consistently utilize supplementary information as constraints. For instance, Miyazaki et al. [1] constrained the ambiguity with the assumption of surface convexity, Mahmoud et al. [2] combined SfP method with shape from shading technology, and Smith et al. [3] expressed polarization and shading constraints as linear equations for estimating surface height. Some studies tried to constrain the ambiguity by limiting reflection type and illumination condition, and others used other techniques, such as photometric stereo, multi-view, and depth, to provide reliable bases for ambiguity resolution. To solve the ambiguity problem with a data-driven method, Ba et al. [4] first introduced deep learning in object-level SfP, and Lei et al. [5] adopted a convolutional neural network with multi-head self-attention to solve the more complicated ambiguity problem in scene-level SfP.

However, achieving high-precision three-dimensional (3D) reconstruction of both global structures and local details is still very challenging due to the following reasons: (1) It is difficult for the physics-based methods to solve the problems in SfP effectively, resulting in a large error in the normal solution results. Besides, the solution process is too complicated when multiple reflection types exist at the same time. (2) Most data-driven methods merely used convolutional neural network (CNN) networks but ignored the clues provided by the global perception, resulting in large local ambiguity and the convex/concave ambiguity in the global structure of estimated results, as shown in Figure 1(b) [6]. (3) Lei et al. [5] introduced global cues into the network to solve local ambiguity with an independent modeling of global and local representations. However, it has been demonstrated that the local and global feature representations should be dependently modeled to corporately provide a better interpretation of images, since the independent modeling limits the power of extracted features and ignores inherent relations between convolution and self-attention. This results in the imperfection to effectively combine advantages of global cues and local features, which makes it hard to solve local convex/concave errors and reconstruct detail texture with high accuracy, as shown in Figure 1(c).

To address the above issues, we propose a new deep learning network, named multi-receptive field interaction network (MRFINet) (as shown in Figure 1(f)), for estimating surface normal based on the SfP method. The key of MRFINet is the interaction of multi-receptive field, i.e., the global receptive field and the local receptive field. The global receptive field allows our proposed network to grasp the surface normal along the boundary and the convexity of the whole object, as well as providing neighborhood surface shape as reference information for solving local ambiguity. The local receptive field can extract subtle convexity change features, which helps to outline surface details and texture. With both global and local receptive fields, the proposed network can achieve high precision estimation of surface normal, as shown in Figure 1(d).

MRFINet. In order to grasp and interact with the multi-receptive field, the proposed MRFINet adopts Conformer as an encoder to constantly extract both global and local features in an interactive way. Besides, with global con-

* Corresponding author (email: xin.tian@whu.edu.cn)
† Peng Y N and Liu R have the same contribution to this work.

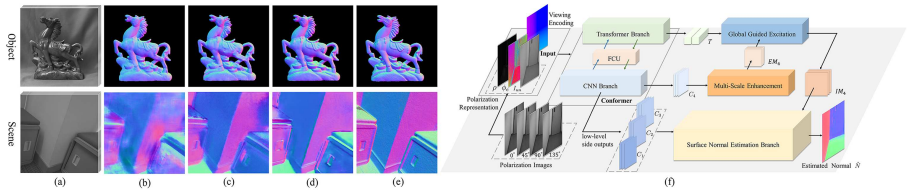

Figure 1 (Color online) Examples of our estimation. (a) The input polarization images; the estimated results of two learning-based SfP methods: (b) Kondo et al. [6] and (c) SPW; (d) the estimated result of MRFINet, which can achieve high-quality reconstruction at both object level and scene level; (e) the ground truth; (f) the overall framework of the proposed MRFINet.

text from the Transformer branch of the Conformer, we utilize the global guided excitation (GGE) module to enhance meaningful features for estimating surface normal. In addition, a multi-scale enhancement (MSCE) module is employed to retain more valuable cues with multi-receptive field interactive information. And the surface normal is estimated from a surface normal estimation (SNE) branch, with the supplementary features of different scales extracted by Conformer's CNN branch.

Encoder. The global context information contains the convexity of the whole object and neighborhood surface shape, which can help resolve the local ambiguities in polarization cues. The local feature can provide subtle convexity change features and contributes to reconstructing regions with fine details. Therefore, to achieve feature extraction of both context information and feature details in an interactive way, we adopt Conformer as the encoder, which consists of a CNN branch and a Transformer branch, adopting a parallel structure.

GGE module. Considering that the CNN high-level features from Conformer contain rich semantic features, we employ a GGE module to use the global clues in the final output from the Transformer branch of Conformer. This design helps guide the expression of high-level semantic features, so as to selectively enhance meaningful features for estimating surface normal, while compressing unconsidered features.

MSCE module. As the final result obtained by the encoder, the high-level CNN feature from Conformer possesses the characteristics of low resolution and a large number of channels. To reduce the information loss in dimension reduction of the high-level feature and retain more valuable cues for surface reconstruction, we introduce an MSCE module, which can inject various multi-scale context information into the original branch and improve the performance of the decoding part.

SNE branch. The high-level CNN features from Conformer have stronger semantic information but also lose the ability to perceive details. However, for high-precision surface reconstruction tasks, fine-grained information (such as texture, edges, and corners) of the image is very important. Therefore, we utilize an SNE branch as the decoder and try to use side outputs of Conformer's CNN branch in different scales as Appendixes A–D to feed into the decoder along with the higher-level features.

Loss function. We supervise both the encoder and decoder during training. For the encoder, we use the cross-entropy loss L_{CE}, which enables the network to learn richer semantic information. While for the decoder, we adopt a cosine similarity loss l_{cosine} on the final output [4]:

$$l_{\mathrm{cosine}} = \frac{1}{W \times H} \sum_{i}^{W} \sum_{j}^{H} \left(1 - \left\langle \hat{N}_{ij}, N_{ij} \right\rangle\right), \qquad (1)$$

where $\hat{N}_{ij}$ and N_{ij} are the estimated result $\hat{N}$ and the ground truth at the pixel location (i, j), respectively. And $\langle \cdot, \cdot \rangle$ denotes the dot product. Constant α is set as 0.1 and the total loss function can be expressed as

$$L = \alpha L_{\mathrm{CE}} + l_{\mathrm{cosine}}. \qquad (2)$$

Experiments. Experiments were performed on an object-level dataset DeepSfP [4] and a scene-level dataset SPW [5]. Our model is implemented on PyTorch and trained on an NVIDIA GeForce RTX 3090 Ti GPU (24 GB). The Conformer-S model used in MRFINet is initialized with a model pre-trained on ImageNet. We train our model for 1000 epochs with a batchsize of 8, use the Adam optimizer with an initial learning rate of 1E−3 and adopt a cosine decay scheduler. We crop images to 512×512 patches in each iteration for memory usage reduction and data augmentation. Some results are illustrated in Figure 1. Our model achieved the best results compared with other approaches under the same conditions. As shown in Appendixes A–D, the theory and background of SfP, algorithm details, experiment analysis, and limitations are described, respectively.

Conclusion. We propose a novel network structure, termed MRFINet, with the integration of global and local receptive fields, to grasp regional reference shape information as well as extract subtle convexity change features, which helps to solve the ambiguity problem in SfP and achieves high-quality reconstruction of surface texture details. Experimental results demonstrate that MRFINet significantly outperforms the existing SfP methods on both the object-level DeepSfP dataset and scene-level SPW dataset.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62471338, 61971315, 62371350).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Miyazaki D, Tan R T, Hara K, et al. Polarization-based inverse rendering from a single view. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2003. 982–987
2 Mahmoud A H, El-Melegy M T, Farag A A. Direct method for shape recovery from polarization and shading. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), 2012. 1769–1772
3 Smith W A P, Ramamoorthi R, Tozza S. Height-from-polarisation with unknown lighting or albedo. IEEE Trans Pattern Anal Mach Intell, 2019, 41: 2875–2888
4 Ba Y, Gilbert A, Wang F, et al. Deep shape from polarization. In: Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, 2020. 554–571
5 Lei C, Qi C, Xie J, et al. Shape from polarization for complex scenes in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 12632–12641
6 Kondo Y, Ono T, Sun L, et al. Accurate polarimetric BRDF for real polarization scene rendering. In: Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, 2020. 220–236