2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Weakly-Supervised Depth Completion during Robotic Micromanipulation from a Monocular Microscopic Image

Han Yang[1,†], Yufei Jin[1,†], Guanqiao Shan[2], Yibin Wang[1], Yongbin Zheng[1], Jiangfan Yu[1], Yu Sun[2,*], and Zhouran Zhang[1,*]

***Abstract*— Obtaining three-dimensional information, especially the z-axis depth information, is crucial for robotic micromanipulation. Due to the unavailability of depth sensors such as lidars in micromanipulation setups, traditional depth acquisition methods such as depth from focus or depth from defocus directly infer depth from microscopic images and suffer from poor resolution. Alternatively, micromanipulation tasks obtain accurate depth information by detecting the contact between an end-effector and an object (e.g., a cell). Despite its high accuracy, only sparse depth data can be obtained due to its low efficiency. This paper aims to address the challenge of acquiring dense depth information during robotic cell micromanipulation. A weakly-supervised depth completion network is proposed to take cell images and sparse depth data obtained by contact detection as input to generate a dense depth map. A two-stage data augmentation method is proposed to augment the sparse depth data, and the depth map is optimized by a network refinement method. The experimental results show that the MAE value of the depth prediction error is less than 0.3 μm, which proves the accuracy and effectiveness of the method. This deep learning network pipeline can be seamlessly integrated with the robotic micromanipulation tasks to provide accurate depth information.**

Index Terms* —— *Biological Cell Manipulation, Automation at Micro/Nano Scales, Deep Learning, Depth Completion

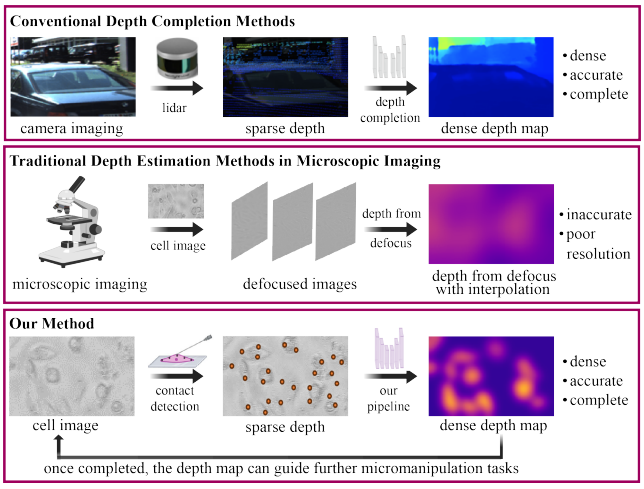

Fig. 1. **Methods for obtaining dense depth maps.** Conventionally, depth sensors such as lidars are used to obtain sparse depth, which is then converted into dense depth map via depth completion. However, in micromanipulation setups, depth sensors are unavailable. Traditional methods, such as depth from focus/defocus, yield depth maps with poor resolution. Our approach employs contact detection within a robotic micromanipulation system, coupled with deep learning methods, to generate dense depth maps.

I. INTRODUCTION

Robotic micromanipulation of biological cells plays an important role in biology and medicine, where individual cells are manipulated by robot end-effectors. Obtaining 3D information, especially z-axis depth information, is crucial to improve manipulation accuracy and efficiency. For instance, a depth map can drive various visual servoing tasks such as microinjection [1], cell morphology measurement [2], path planning for cell transportation [3], 3D reconstruction of cellular scenes [4], and virtual reality applications [5].

Cell manipulation is usually performed under microscopes, where depth sensors such as lidars or RGB-D cameras cannot be integrated (Fig. 1); hence, traditional methods rely on vision algorithms to infer depth from microscopic images. For instance, depth from focus and depth from defocus [6] [7] [8] compute depth based on the sharpness or blurring of image pixels. They both suffer from the low axial resolution of optical microscopes. Objects within the depth of field are simultaneously in focus, thus resulting in poor resolution in the depth map (Fig. 1). Different imaging modalities such as holographic and confocal microscopy have also been used to obtain depth information [9] [10]. However, holographic microscopy has slow imaging speeds and complex digital processing, making it unsuitable for real-time or high-speed imaging [11], whereas confocal microscopy requires the cells to be stained [12] [13]. Although additional side-view cameras can provide 3D stereo vision [14], these cameras cannot view through the cell monolayers where cells are occluded by each other in side-view images. These methods are not suitable for providing accurate depth information during robotic cell manipulation.

To obtain accurate depth information for micromanipulation, contact detection has been introduced in the past decades [15], [16], [17], [18]. It lowers the robot end-effector that is built within the micromanipulation system along the z-axis to touch the cell, and once cell deformation is detected, its corresponding z position can be obtained. The obtained depth information can then be seamlessly used for subsequent cell manipulation tasks. This method can bypass the axial resolution limitation of microscopes, because the cell deformation "amplifies" the number of detectable pixels for depth measurement; hence, the achieved resolution is determined by the resolution of robotic micromanipulators

† Equal contributions. *Correspondence to zhangzhuoran@cuhk.edu.cn, sun@mie.utoronto.ca
[1] School of Science and Engineering, the Chinese University of Hong Kong, Shenzhen, China.
[2] Department of Mechanical and Industrial Engineering, University of Toronto, Canada

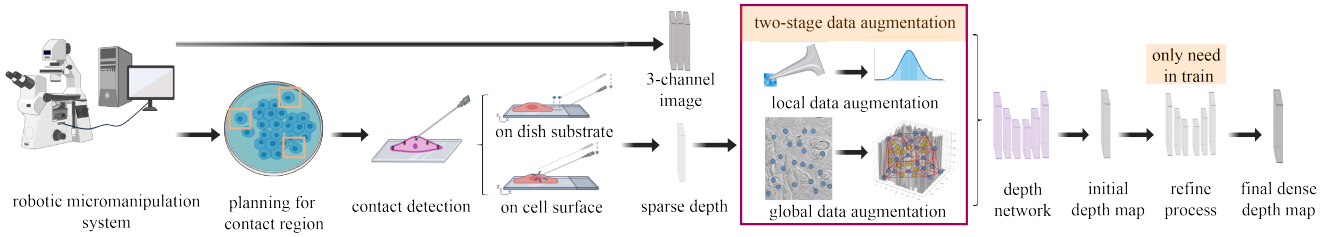

Fig. 2. **Pipepine for depth completion during robotic micromanipulation.** The pipeline first plans regions for contact detection, then in each region automated contact detection is performed only once to avoid repeated experiments on regions with similar image features. The collected sparse depth data are then augmented and fed into a depth completion network, followed by a refinement process to generate a dense depth map.

(e.g., 0.1 μm [16]), making it a promising candidate method to acquire accurate depth values.

The low efficiency of the contact detection method, however, makes it only suitable to acquire sparse depth data, which requires a depth completion step to further construct a dense depth map for the entire image. The extremely low sparsity is a major challenge. Ideally, each pixel in an image should correspond to a depth value, and a 640×480-pixel image requires ∼300,000 contacts. For conventional depth completion tasks such as depth completion of sparse lidar data, a sparsity of 5% is considered as extremely low sparsity. However, this still requires 15,000 contacts on a single image. A few hundreds of contacts result in a sparsity of less than 1%. For training existing depth estimation models such as [19] [20], supervisory information for such low sparsity is not sufficient for the models to converge. Although sparse depth information can be used as weakly supervised information to obtain dense depth information using deep learning techniques [21] [22] [23], it remains unknown whether such low sparsity could be used as weekly supervision to train deep depth completion models. In addition, depth completion has not been attempted on microscopic images, which contain fewer image features compared to images in indoor or outdoor RGB-D datasets [24] [25].

This paper aims to develop a deep learning pipeline to generate dense depth map for use in robotic micromanipulation process, which, to the best of our knowledge, is the first deep learning process for acquiring dense depth information during robotic micromanipulation. Our technique utilizes the commonly used contact detection step to obtain sparse depth values and does not change the flow of manipulation. The pipeline takes the cell image and sparse depth values as input. To enable model training with extremely sparse depth data, a two-stage data augmentation method is proposed. A depth completion network is designed, followed with a lightweight depth refinement network to fully utilize the limited amount of depth data. Once completed, the dense depth map provides guidance for subsequent micromanipulation tasks. Our contributions are as follows:

- As a proof-of-concept attempt, this work achieved deep learning-based depth completion on microscopic images. The proposed technique uses contact detection, which is a common step during micromanipulation, to obtain sparse depth data and can be seamlessly integrated with micromanipulation processes. The technique lays foundation for further improvement of depth estimation methods in robotic micromanipulation.
- A weakly-supervised deep learning pipeline is proposed where a lightweight depth completion model is trained and refined to reconstruct dense depth map from sparse depth data. Once completed, the dense depth map can be used to guide subsequent micromanipulation tasks.
- To address the challenge of extremely sparse depth data, a two-stage data augmentation method is developed to enable model training. The method uses both local information near the contacted pixel and global information of the cell scenes to augment the sparse depth data. The effectiveness of the data augmentation method was confirmed using ablation studies.

II. Overview of the Depth Completion Pipeline

As shown in Fig. 2, the depth completion pipeline starts from collecting the input data, including the cell image and sparse depth data obtained by contact detection on both the substrate of the culture dish and on cell surface. By definition in conventional depth completion, each pixel's depth refers to the distance between the camera and object. In micromanipulation, the camera is mounted on a microscope, and the distance between the camera and object (cell) reflects the length of the optical path, which cannot be used directly for micromanipulation. Therefore, depth is represented by each pixel's corresponding height on the cell surface relative to the dish substrate. This convention ensures that the depth information (cell height relative to substrate) can be readily used for micromanipulation tasks.

Before contact detection, a planning method is developed to identify regions with similar image features. Contact detection is then performed on each similar region only once, thus avoiding unnecessary repeated contact detection. Specifically, the image is divided into 16 equal sub-regions to lower complexity and boost computational efficiency. Feature matching within each sub-region is performed using Oriented Fast Rotated BRIEF (ORB) descriptors [26], with key points paired and matched through brute force. The Euclidean distance of matched points offers a similarity metric, identifying the region with the most corresponding features in each image sub-region.

After contact detection micromanipulation, the obtained sparse depth is augmented by a two-stage data augmentation

method. The depth data then serve as weak supervision for the depth completion network to generate an initial depth map, which is then refined with extra depth data to generate the final dense depth map.

All the above operation is performed by a robotic micromanipulation system which consists of an inverted microscope (ECLIPSE Ti2, Nikon Inc.) equipped with a motorized XY-stage (ProScan, Prior Scientific Inc.). This stage provides a range of 75 mm and a resolution of 0.01 μm. A 3-DOF motorized micromanipulator (uMp-285, Sensapex Inc.) is integrated to hold a glass micropipette for contact detection to obtain sparse depth. The micropipette is made of glass tubes with a micropipette puller (P-97, Sutter Inc.) following the method in [17] and has a tip with outer diameter of 500 nm and inner diameter of 300 nm. Visual feedback is supplied by a camera (Basler acA1920-40u Inc.), enabling image-based visual servo control for micropipette tip localization and detection of cell surface deformation. All user interface interactions and deep learning model computations are managed by a computer with a GPU 3070Ti.

III. METHODOLOGY

A. Local & Global Data Augmentation

The depth data collected by contact detection are too sparse (e.g., $\sim$200 depth values for a 640×480 image, corresponding to a sparsity of 0.06%) to train a neural network. Hence, a two-stage data augmentation method is proposed to make it feasible for neural network training. The proposed method first augments depth data near each contact location (local data augmentation), then supplements sparse depth information with 3D scene topology of the global image (global data augmentation).

Local Data Augmentation. Ideally, each pixel in the depth map corresponds to a depth value. However, in contact detection, the depth value is determined by detecting the contact between the micropipette tip and the cell membrane or substrate. Considering that the actual physical size of the micropipette tip is larger than one pixel (Fig.4a), each contact location covers an area instead of a single pixel. Hence, utilizing this nature of contact detection, multiple depth values could be obtained from a single contact, thus augmenting the available depth data. As shown in the example in Fig. 4a), locally near the contact location, a 3×3-pixel area is covered by the pipette tip, and minimal depth variations are expected over such a small area. Hence, the obtained depth value from contact detection is assigned to the center pixel of the 3×3 area. A Gaussian distribution is used to distribute depth values among the area. This method not only solves the problem of depth value assignment near the local contact location but also augments the obtained depth data from each contact by 9 times, thus eliminating the need for multiple contacts at nearby locations.

Global Data Augmentation. Although augmented by 9 times after local data augmentation, the depth data are still sparse. For instance, still considering the example of 200 contacts for a 640×480 image, 1,800 depth values can be obtained after local augmentation, corresponding to a sparsity of 0.54%. For global data augmentation on the entire image as shown in Fig. 4 b), conventional methods directly interpolate the sparse depth data. However, depth reflects information in three-dimensional (3D) space, and direct interpolation of the depth data misses the information of 3D scene topology [27], [28]. For instance, with depth information, the entire cell surface can be regarded as a 3D scene, and each nearby region of the surface can be regarded as a triangular plane in 3D space. In this work, the triangular planes are recovered as constraints for interpolating the sparse depth data in 3D. The objective of this step is to obtain a preliminary depth map for global augmentation of the whole depth map.

To incorporate 3D scene information, the lifting transformation [29] is first employed to map sparse depth from 2D space to 3D space. Let $d_s \in \mathbf{R}^{n\times 2}$ represent the set of points in the 2D space and d'_s represent the points after lifting. The lifting transformation f is represented by:

$$d'_s = f(d_s) \in \mathbf{R}^{n\times 3} \tag{1}$$

Next, the convex hull of depth information is computed [30], yielding the Delaunay triangulation of the cell scene in Barycentric coordinates. Within each triangle, the sparse depth data are augmented through bilinear interpolation to provide initial dense depth data. Let T represent a triangular mesh and $b(T)$ its representation in Barycentric coordinates, then the interpolated mesh T' is computed as:

$$T' = \sum_{i=1}^{3} b_i(T)(d'_s)_i \tag{2}$$

Projecting the generated triangular skeleton and initial dense depth data back onto the 2D plane provides a more detailed approximation of the scene. The augmented depth data will be further refined in the following depth network to get the final dense depth map.

B. Depth Completion Network & Refinement

In order to complement the sparse depth information, a dual-branch encoder network is developed. The coarse sparse depth information obtained after two-stage data augmentation and the corresponding cell images are used as inputs, respectively, and the cell images guide the depth completion task. The network obtains a dense depth map by integrating the local-to-global information. The validated final depth map provides a priori information for subsequent micromanipulation tasks.

Network Architecture. In the design of the depth network, a dual Encoder architecture is employed, drawing inspiration from the late-fusion strategy [31]. The architecture is to handle depth and image data differently, reflecting distinct characteristics of these data types.

For depth data, the network uses a large kernel size (7×7, 5×5, and 5×5) for 3 convolution layers. This decision is guided by an attempt to reduce network complexity while capturing the spatial structure inherent in the depth data. In contrast, image data that contain texture information are

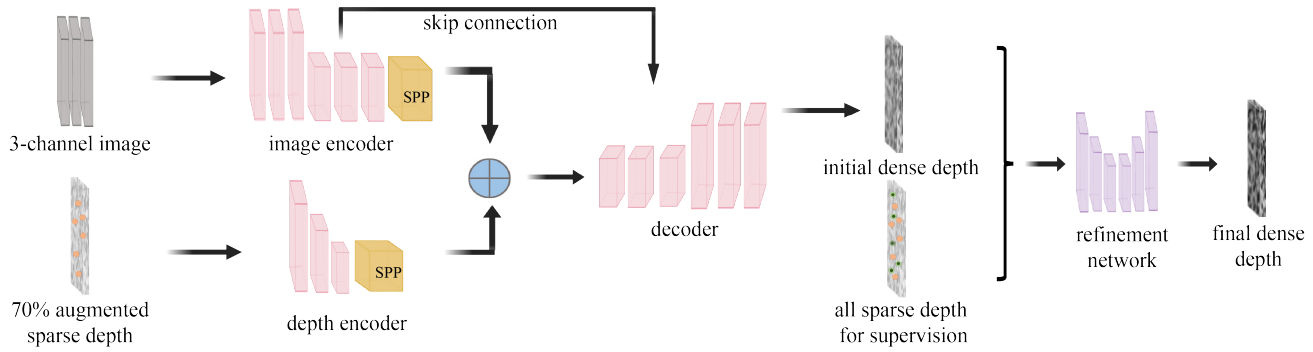

Fig. 3. **Architecture for the depth completion network**. The network consists of a dual-encoder architecture leveraging Spatial Pyramid Pooling (SPP) blocks to extract local-to-global context from depth and images. The decoded initial depth map is refined via a Non-local Spatial Propagation Network (NSPN) to yield the final dense depth map.

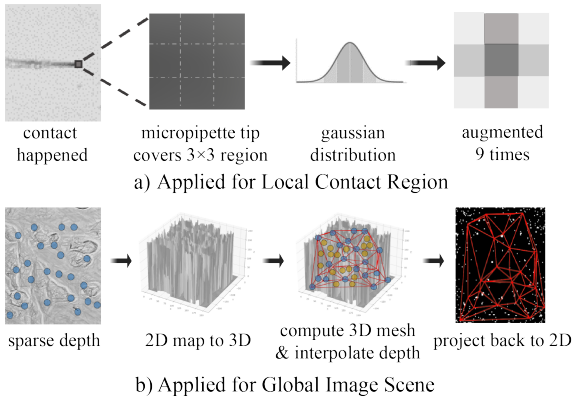

Fig. 4. **Two-stage Data Augmentation.** a) Local data augmentation uses Gaussian distribution to augment depth data for pixels near the contact location. b) Global data augmentation incorporates 3D scene topology of the entire cell image for data augmentation.

processed with a smaller kernel size (3×3 kernel size) for 6 convolution layers. This allows the network to learn the image information thoroughly, which subsequently aids the depth completion process.

In order to save the computational cost of the network and to avoid duplicate feature extraction, both the depth Encoder and the image Encoder incorporate spatial pyramid pooling (SPP) blocks [32]. These SPP blocks enable the network to learn contextual information at various scales, from local details to global scene structure. Additionally, each Encoder is enhanced with residual blocks [33] and deformable convolutions [34]. Deformable convolutions perform position-variable convolution operations and enable the convolution kernel to adaptively adjust its shape and position by learning spatial transformation parameters, which further makes it more adaptable to the sparsity of depth data. In addition, due to the nonlinear nature of the convolution operation, deformable convolution can effectively capture complex nonlinear relationships in sparse depth.

The contextual information learned from each branch is then concatenated, creating a feature map that encapsulates both depth and image features. These combined features are then propagated through the decoder, which employs a series of deconvolution layers with batch normalization for feature reconstruction. The output from different layers of the decoder is upsampled through interpolation to match the original input resolution.

Refinement. To fully utilize the limited amount of depth data, a train-refinement strategy is used to improve the accuracy of the network. The collected depth data by contact detection are first divided into two parts: 70% and 30%. Among them, 70% of data are used as initial supervision to guide the prediction of an initial dense depth map (the depth map directly from the network output). The remaining 30% of data are used as additional supervision to further optimize and refine the initial depth map. This refinement process is applied only in model training without the need for additional depth information during model inference, thus allowing for seamless integration into the robot micromanipulation.

The 30% additional supervision data, together with the initial dense depth map, are refined by spatial propagation networks (SPN) [35] to optimize data uncertainty handling and enhance object surface continuity. Considering that the affinity learning of SPN is limited to the local pixels, we employ a non-local spatial propagation network (NSPN) [36] and learnable affinity normalization to suppress the effect of unreliable depth values. Experiments demonstrated that the NSPN effectively improved model accuracy.

Loss Function. The training loss function L_{train} comprises two terms: the similarity between the predicted and the sparse depth values L_{sparse}, and a term to ensure the smoothness of the predicted results L_{smooth}.

$$L_{train} = \lambda L_{sparse} + \beta L_{smooth} \tag{3}$$

where λ and β are the associated weights respectively.

In the depth-completion task, the main goal is to minimize the difference between the predicted depth value $\hat{d}(x)$ and the ground truth depth value $d(x)$ obtained by contact detection. We define this sparse depth consistency as L_{sparse}:

$$L_{sparse} = \frac{1}{|GT_s|} \sum_{x \in GT_s} |\hat{d}(x) - d(x)| \tag{4}$$

In addition, a smoothness loss function L_{smooth} is introduced to solve the problem of blurry boundaries in the depth map. Specifically, the L1 criterion [37] is applied to the

gradient of the predicted depth map in both x and y directions to ensure boundary clarity and depth map smoothness during depth completion.

$$L_{\text{smooth}} = \frac{1}{|GT_s|} \sum_{x \in GT_s} \left(E_X |\nabla_X \hat{d}(x)| + E_Y |\nabla_Y \hat{d}(x)| \right) \tag{5}$$

where E_X and E_Y are the edge-awareness weights. This addition of the smoothness supervision term helps to maintain sharp edges and fine details in the completed depth map.

IV. RESULTS AND DISCUSSION

In experiments, depth values were collected on Hela human cancer cells using the contact detection method as described in [16]. The cells were cultured in DMEM medium (OriCell Inc.) supplemented with 10% FBS (OriCell Inc.). Prior to experiments, the cells were passaged, seeded into 60 mm Petri dishes, and cultured for 24 hours to form a cell monolayer. The contact detection algorithm has a success rate of more than 95% and a normalized error of less than 4% [16]. The training dataset contains 240 cell images with a resolution of 640 $\times$ 480, together with depth values from $\sim$72,000 contact detection operations.

TABLE I
QUANTITATIVE EVALUATION OF DEPTH COMPLETION RESULTS

Number of Depth Values	MAE	RMSE	iMAE	iRMSE [μm]
300	**0.26**	**0.33**	**0.28**	**0.30**
150	1.48	1.12	1.53	1.26
50	2.97	3.62	3.12	3.41

TABLE II
ABLATION STUDY FOR 300 CONTACT TIMES

Two-stage Data Augmentation		Refinement	RMSE [μm]
Local	Global		
✓	✓	✓	**0.33**
×	✓	✓	0.81
✓	×	✓	NA
✓	✓	×	0.58
×	×	✓	NA
×	×	×	NA

A. Depth Completion Performance

To evaluate network performance, additional sparse depth information was obtained from 60 Hela cell images. These images were not included in the dataset for network training or refinement. For each cell image, contact detection was performed 300 times to obtain ground truth depth values. Qualitative prediction results are shown in Fig. 5. The network successfully predicted depth maps in different Hela cell images. The cell boundary near the dish substrate showed the lowest depth (height) value, whereas the cell center showed a higher depth (height) value, which is in agreement with Hela cells' 3D morphology. In terms of quantitative performance evaluation, standard evaluation metrics including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), inverse Mean Absolute Error (iMAE) as well as inverse Root Mean Square Error (iRMSE) were used. The network showed an MAE of 0.26 μm and a RMSE of 0.33 μm (the first row of Table I). Considering the maximum depth (height) is around 5$\sim$6 μm for each cell, the achieved error of $\sim$0.3 μm provides accurate depth information for micromanipulation tasks.

B. Comparison with Existing Methods

The performance of the developed network was further qualitatively compared with existing depth completion methods, including bilinear interpolation, and the five-point ellipse fitting method by Liu et al. [16]. For the comparison, local or global data augmentation was not used for both of these methods. For all comparisons, contact detection was performed 300 times to provide sparse depth values. Not surprisingly, the available depth values were too sparse for bilinear interpolation, and even the shapes of cells were not recognizable in the depth map [see Fig. 6b)].

The cell five-point ellipse fitting method [16] approximated each cell with an ellipse and the depth values were obtained by contact detection on five locations on each cell, including midpoints of long and short axes of the ellipse and the center of the ellipse. As shown in Fig. 6c), although this assumption yielded sharp boundaries for each cell in the depth map, the fitted depth values had almost identical depth values on each cell, which could not accurately reflect the real-world cell morphology as they might not exactly conform to a regular ellipse shape. In addition, this method can only obtain depth information for five contact points on the cell surface, which was not sufficient to recover a dense depth map of the whole cell. Compared to the above two methods, our method was able to acquire cell boundaries as well as more accurate depth information for the cell surface height. In addition, our network took only 0.76 seconds to generate a depth map of an image and thus can be seamlessly integrated into the micromanipulation process. This improved depth map can be used to more accurately micromanipulations such as cell morphology measurement.

C. Ablation Study

In order to gain insight into the impact of each component on network performance, we further conducted an ablation study of the two-stage data augmentation algorithm and the network refinement step, with results shown in Table II.

Without local data augmentation, the discrepancy between the network prediction and the ground truth increased (RMSE from 0.33 μm to 0.81 μm), suggesting that local data augmentation contributed to improving the accuracy. This algorithm achieved a 9 times augmentation of the data within the 3$\times$3-pixel area covered by the micropipette tip by augmenting the depth values with a Gaussian distribution for each depth value. Without global data augmentation, the network did not converge; hence, RMSE values were not available. Global data augmentation took into account

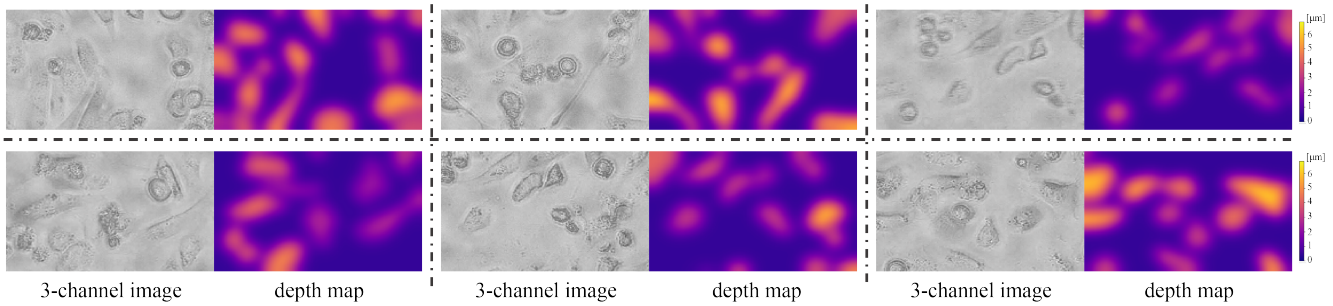

Fig. 5. **Qualitative depth completion results by the proposed method.** The original images of Hela cells and the corresponding depth map are shown in parallel.

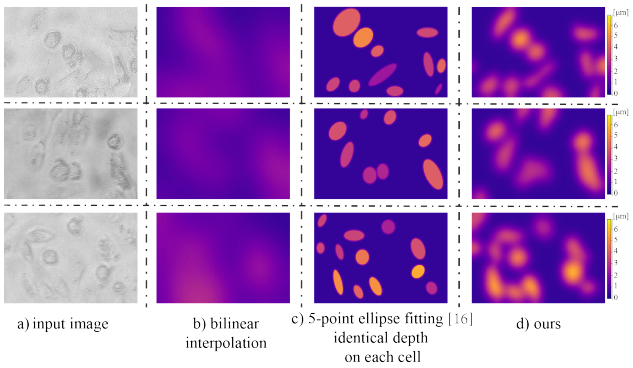

Fig. 6. **Qualitative comparison with existing methods.** a) The original cell image. b) Results of bilinear interpolation of the sparse depth directly. c) The depth map obtained by the 5-point ellipse fitting method [16]. The fitted result yielded almost identical depth on each cell. d) The depth map obtained by our method.

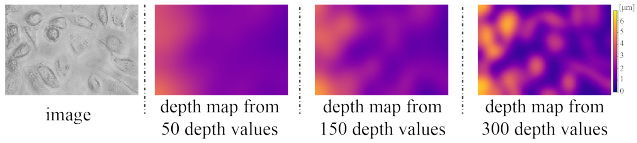

Fig. 7. **Depth completion results from different depth values (contact times)**. The same image was contacted by 50, 150, and 300 times, respectively, and more depth values yield better depth completion results.

the topological constraints of the scene and interpolated the sparse depth values of the entire cell scene for augmentation. This further achieved the dual effect of reducing the zero-depth values while making the augmentation results closer to the true depth distribution. The results highlight the importance of the proposed global data augmentation method for depth completion. Without the refinement step, the prediction error increased from 0.33 μm to 0.58 μm. Although the error did not increase as much as ablating local data augmentation, the results still showed the effectiveness of the refinement strategy. This refinement step optimized the depth value of the depth network prediction error by introducing additional depth data as supervision for training.

D. Effect of Sparsity of Depth Values

We further evaluated the effect of different sparsity of depth samples on model performance. For the same cell image, contact detection was performed 50, 100, and 300 times, respectively. The obtained depth was used for training the depth completion network.

Not surprisingly, denser depth values led to better depth completion results. For qualitative comparisons [see Fig.7], depth maps obtained with only 50 contacts (depth values) showed large errors with cell shapes even unrecognizable. Using 150 contacts yielded a clearer depth map, where the cell boundaries and depth variations were more pronounced. When the number of contacts was increased to 300, the cell contours and depth variations of the depth maps were much clearer. For quantitative evaluation, four metrics were used, including MAE, RMSE, iMAE, and iRMSE (Table I). The errors for all performance metrics increased as the number of contacts decreased. Using 300 contacts led to the smallest errors for all four metrics (0.26, 0.33, 0.28, and 0.30 μm, respectively), yielding the effectiveness of the proposed depth completion approach.

V. CONCLUSION

This work presents a deep learning pipeline for generating dense depth maps from sparse depth data during micromanipulation processes. Different from conventional depth sensors such as lidar, robotic micromanipulation obtains depth information by physically detecting the contact between the end-effector and the manipulated object (e.g., a cell). To utilize the extremely sparse depth data obtained by contact detection, this work proposes a two-stage data augmentation approach, using both local and global information of the image. A depth completion network is developed to fill the sparse depth, and a network refinement technique is applied to fully utilize the limited amount of depth data to achieve more precise dense depth maps. The presented depth completion pipeline can be seamlessly integrated into micromanipulation tasks. As a proof-of-concept for depth completion using microscopic imaging, the technique also lays foundation for further improvement of depth estimation methods in robotic micromanipulation tasks.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China (2023YFE0205500), in part by the National Natural Science Foundation of China (62203374), in part by Guangdong Basic and Applied Basic Research Foundation (2021A1515110023), and in part by Shenzhen Science and Technology Program (RCBS20210706092254072).

REFERENCES

[1] W. Wang, X. Liu, D. Gelinas, B. Ciruna, and Y. Sun, "A fully automated robotic system for microinjection of zebrafish embryos," *PloS one*, vol. 2, no. 9, p. e862, 2007.

[2] C. Dai, Z. Zhang, J. Huang, X. Wang, C. Ru, H. Pu, S. Xie, J. Zhang, S. Moskovtsev, C. Librach, *et al.*, "Automated non-invasive measurement of single sperm's motility and morphology," *IEEE Transactions on Medical Imaging*, vol. 37, no. 10, pp. 2257–2265, 2018.

[3] V. Venkatesan and D. J. Cappelleri, "Path planning and micromanipulation using a learned model," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3089–3096, 2018.

[4] Y. Wang, X. Ji, and Q. Dai, "Key technologies of light field capture for 3d reconstruction in microscopic scene," *Science China Information Sciences*, vol. 53, pp. 1917–1930, 2010.

[5] K.-C. Kwon, K. H. Kwon, M.-U. Erdenebat, Y.-L. Piao, Y.-T. Lim, Y. Zhao, M. Y. Kim, and N. Kim, "Advanced three-dimensional visualization system for an integral imaging microscope using a fully convolutional depth estimation network," *IEEE Photonics Journal*, vol. 12, no. 4, pp. 1–14, 2020.

[6] M. Shang, T. Kuang, H. Zhou, and F. Yu, "Monocular microscopic image 3d reconstruction algorithm based on depth from defocus with adaptive window selection," in *2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2. IEEE, 2020, pp. 17–21.

[7] Y. Matsubara, K. Shirai, Y. Ito, and K. Tanaka, "Pixel-wise parallel calculation for depth from focus with adaptive focus measure," *Multidimensional Systems and Signal Processing*, vol. 33, no. 1, pp. 121–142, 2022.

[8] N. Madali, A. Gilles, P. Gioia, and L. Morin, "Automatic depth map retrieval from digital holograms using a depth-from-focus approach," *Applied optics*, vol. 62, no. 10, pp. D77–D89, 2023.

[9] A. Sahu, O. Yélamos, N. Iftimia, M. Cordova, C. Alessi-Fox, M. Gill, G. Maguluri, S. W. Dusza, C. Navarrete-Dechent, S. González, *et al.*, "Evaluation of a combined reflectance confocal microscopy–optical coherence tomography device for detection and depth assessment of basal cell carcinoma," *JAMA dermatology*, vol. 154, no. 10, pp. 1175–1183, 2018.

[10] M. Panahi, R. Jamali, V. F. Rad, M. Khorasani, A. Darudi, and A.-R. Moradi, "3d monitoring of the surface slippage effect on micro-particle sedimentation by digital holographic microscopy," *Scientific Reports*, vol. 11, no. 1, p. 12916, 2021.

[11] C. Martin, L. E. Altman, S. Rawat, A. Wang, D. G. Grier, and V. N. Manoharan, "In-line holographic microscopy with model-based analysis," *Nature Reviews Methods Primers*, vol. 2, no. 1, p. 83, 2022.

[12] D. S. Gareau, "Feasibility of digitally stained multimodal confocal mosaics to simulate histopathology," *Journal of biomedical optics*, vol. 14, no. 3, pp. 034 050–034 050, 2009.

[13] S. González and Z. Tannous, "Real-time, in vivo confocal reflectance microscopy of basal cell carcinoma," *Journal of the American Academy of Dermatology*, vol. 47, no. 6, pp. 869–874, 2002.

[14] Z. Zhang, X. Wang, J. Liu, C. Dai, and Y. Sun, "Robotic micromanipulation: Fundamentals and applications," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 181–203, 2019.

[15] W. Wang, X. Liu, and Y. Sun, "Contact detection in microrobotic manipulation," *The International Journal of Robotics Research*, vol. 26, no. 8, pp. 821–828, 2007.

[16] J. Liu, Z. Zhang, X. Wang, H. Liu, Q. Zhao, C. Zhou, M. Tan, H. Pu, S. Xie, and Y. Sun, "Automated robotic measurement of 3-d cell morphologies," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 499–505, 2016.

[17] J. Liu, V. Siragam, Z. Gong, J. Chen, M. D. Fridman, C. Leung, Z. Lu, C. Ru, S. Xie, J. Luo, *et al.*, "Robotic adherent cell injection for characterizing cell–cell communication," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 1, pp. 119–125, 2014.

[18] J. Liu, C. Shi, J. Wen, D. Pyne, H. Liu, C. Ru, J. Luo, S. Xie, and Y. Sun, "Automated vitrification of embryos: A robotics approach," *IEEE Robotics & Automation Magazine*, vol. 22, no. 2, pp. 33–40, 2015.

[19] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[20] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.

[21] J. Gu, Z. Xiang, Y. Ye, and L. Wang, "Denselidar: A real-time pseudo dense depth guided depth completion network," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1808–1815, 2021.

[22] Y. Zhao, L. Bai, Z. Zhang, and X. Huang, "A surface geometry model for lidar depth completion," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4457–4464, 2021.

[23] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the cpu," in *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 2018, pp. 16–22.

[24] Z. Cai, J. Han, L. Liu, and L. Shao, "Rgb-d datasets using microsoft kinect or similar sensors: a survey," *Multimedia Tools and Applications*, vol. 76, pp. 4313–4355, 2017.

[25] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[26] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[27] G. Graber, T. Pock, and H. Bischof, "Online 3d reconstruction using convex optimization," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2011, pp. 708–711.

[28] A. Wong and S. Soatto, "Unsupervised depth completion with calibrated backprojection layers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 747–12 756.

[29] K. Q. Brown, "Voronoi diagrams from convex hulls," *Information processing letters*, vol. 9, no. 5, pp. 223–228, 1979.

[30] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software (TOMS)*, vol. 22, no. 4, pp. 469–483, 1996.

[31] M. Jaritz, R. De Charette, E. Wirbel, X. Perrotton, and F. Nashashibi, "Sparse and dense data with cnns: Depth completion and semantic segmentation," in *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 52–60.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[33] K. H. et al., "Identity mappings in deep residual networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 630–645.

[34] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[35] S. Liu, S. De Mello, J. Gu, G. Zhong, M.-H. Yang, and J. Kautz, "Learning affinity via spatial propagation networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[36] Y. Lin, T. Cheng, Q. Zhong, W. Zhou, and H. Yang, "Dynamic spatial propagation network for depth completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1638–1646.

[37] A. Wong, X. Fei, S. Tsuei, and S. Soatto, "Unsupervised depth completion from visual inertial odometry," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1899–1906, 2020.