Data-Driven Microscopic Pose and Depth Estimation for Optical Microrobot Manipulation

Dandan Zhang,^{*,†} Frank P.-W. Lo,[†] Jianqing Zheng,[‡] Wenjia Bai,[†]

Guang-Zhong Yang,^{*,¶} and Benny Lo[†]

[†]*Imperial College London, London, UK*

[‡]*Oxford University, Oxford, UK*

[¶]*Shanghai Jiao tong University, Shanghai, China*

E-mail: d.zhang17@imperial.ac.uk; gzyang@sjtu.edu.cn

Phone: +44 07547525412

Abstract

Optical microrobots have a wide range of applications in biomedical research for both in *vitro* and in *vivo* studies. In most microrobotic systems, the video captured by a monocular camera is the only way for capturing the movements of microrobots, and only planar motion can be captured by a monocular camera system.

Accurate depth estimation is essential for the 3D reconstruction or autofocus of micro-platforms, while the pose and depth estimation are necessary to enhance the 3D perception of the micro-robotic systems to enable dexterous micro-manipulation and other tasks. In this paper, we propose a data-driven method for pose and depth estimation in an optically manipulated microrobotic system. Focus measurement is used to obtain features for Gaussian Process Regression (GPR), which enables precise depth estimation. For mobile microrobots with varying poses, a novel method is developed based on a deep residual neural network with the incorporation of prior domain knowledge about the optical microrobots encoded via GPR. The method can simultaneously

track microrobots with complex shapes and estimate the pose and depth values of the optical microrobots.

Cross-validation has been conducted to demonstrate the sub-micron accuracy of the proposed method and precise pose and depth perception for microrobots. We further demonstrate the generalisability of the method by adapting it to microrobots of different shapes using transfer learning with few-shot calibration. Intuitive visualization is provided to facilitate effective human-robot interaction during micro-manipulation based on pose and depth estimation results.

Keywords

Optical Microrobot, Pose Estimation, Deep Learning

1 Introduction

Optical microrobots represent a growing area of robotics research for biomedicine,¹ owing to advances in materials and micro-fabrication technology. However, due to the inherent challenges of micro-fabrication, assembly and functionalization of sensors in micro-scale, *ex vivo* and *in vitro* microrobotic experimental setups normally rely on the camera view as their main sensing modality. Therefore, the vision system is essential for simultaneous, independent trapping and manipulation of different micro-objects. The camera views can be used to implement computer vision algorithms for microrobot monitoring, which provides a form of feedback to improve user perception and development of closed-loop control techniques. To this end, the perception of microrobots via microscopic view is worth exploring.

For the perception of micro-particles which do not have high requirements on pose estimation, position tracking has been widely explored. For example, position tracking of micro particle has been explored for the digital video microscopy,² which has been used for optical tweezer calibration,³ biomolecular forces measurement,⁴ exploration of rheology of complex

fluids,⁵ and assessment of mechanical properties of soft tissues.⁶ Some of the existing algorithms have been developed to track micro particles, especially for tracking fluorescent particles and molecules.^{7,8} Machine-learning techniques have been explored for fast and accurate feature localization in holograms of colloidal particles,⁹ while convolutional neural networks have shown their capabilities in particle tracking for microscopy applications.^{10–12} However, for microrobots with complex shapes, robust vision-based techniques for tracking microrobots and 6D pose estimation need further development.

In addition to tracking micro-particles, the tracking and 6D pose estimation of microrobots are significant. Accurate depth and pose estimation plays an important role in closed-loop control of microrobots and improving the microscale 3D perception for the operator. It can support automatic micro-assembly tasks and indirect manipulation of objects such as cells.¹³ For large-scale robotic systems, depth can be estimated by using time-of-flight camera, structured illumination, pulsed Lidar or stereo camera technologies. However, for micro-scale robotic systems, most of the available sensing and camera technologies are not able to estimate depth reliably. The transparency of optical microrobots and varying illumination levels set by the users bring additional challenges to depth estimation. Compared to less transparent or opaque objects, a transparent object can complicate effective pose estimation, since the motion of the microrobots along the optical axis may be unobservable due to parallel projection.¹⁴

Previous work in the area has proposed a number of methods for depth or pose perception of microrobots. For example, the depth recovery of microgrippers and Carbon Nanotubes (CNT) for automation applications has been proposed.¹⁵ In microrobotics research, feature based depth estimation for micro-grippers/micro-robots with Scanning Electron Microscope (SEM)^{16–18} and 3D orientation estimation of the microrobot have been investigated.¹⁹ Model-based tracking of magnetic intraocular microrobot has also been proposed.²⁰ However, existing methods are often designed for a specific experimental setup and they cannot be adapted to different micro-robotic systems, such as mobile microrobots with complex shape and have

various poses.

Compared to traditional approaches, machine learning based algorithms, such as neural networks, can provide more general solutions to assist the processing of optical images for clinical and biomedical applications.²¹⁻²³ For optical microrobot depth estimation, convolutional neural networks (CNNs) combined with long short-term memory (LSTM) have been used²⁴ for sequential data regression based on gray-scale monocular images generated in an Optical Tweezer(OT) setup during offline analysis, where trajectory reconstruction of microrobots was implemented following the periodic trajectory (such as the sinusoidal trajectory and the triangular trajectory). However, the movements of the microrobot are often random during micro-manipulation processes. The model trained on one group of pre-defined periodic trajectories may not work well in others, since the microrobots can move with random motions. Moreover, the experimental results of prior works were only validated on one model, and it has not been tested for different microrobots with complex shapes.

In addition to depth estimation, 3D pose estimation in microrobotics could also benefit from the advances in artificial intelligence. For example, CNNs have been employed for the estimation of 3D pose of an object from a single image.²⁵ A CNN-based method for estimating the 3D pose and depth of optically transparent microrobots has been developed.²⁶ However, for a neural network to be able to estimate the 3D pose of different microrobots, a large database should be collected, which consists of microscopic images of different microrobots with different shapes. The network for pose and depth estimation has to be trained from the raw data of the different robots, which is time-consuming. Moreover, the pose estimation was implemented with a relative orientation estimation mode, which means that accumulative errors may severely affect the accuracy of pose estimation. Therefore, a more generic method is needed for pose and depth estimation for other microrobots given that the available data for training the machine learning model is limited. It has been proved that transfer learning framework can assist the discovery of the general underlying physical rules.²⁷ Therefore, to enable the method proposed in this paper to be adaptable to multiple micro-robotic systems

with various designs, transfer learning is used to reduce the need of collecting a large database for each microrobot.

To realize the target of precise pose and depth estimation mentioned above, an algorithm is designed in this paper. With accurate pose and depth estimation of the optical micro-platform/microrobots based on the video data, additional views can be used to enhance the 3D perception capabilities of the operators in the micro-scale. Since the camera can only capture the top-to-down view, an additional side view or 3D view can enable the operator to have a better sense of the microrobot's depth and pose within the targeted workspace. This paves a way for the automation in the 3D space for micro-robotic system, which has requirement on the sensing capabilities.

The main contributions of this paper are listed as follows. 1) Depth estimation for optical micro-platform is constructed based on focus measurement and GPR model. 2) A Hierarchical Semi Supervised ResNet architecture is developed in a multi-task learning manner, which can be used for optical microrobots pose and depth estimation. 3) A GPR-ResNet Hybrid model is developed to further improve the accuracy of depth estimation of microrobot with various poses, which constructs a more precise depth estimation approach for optical microrobots. 4) Domain adaption is verified to demonstrate the generality of the proposed method for applications in different optical microrobotic systems. 5) A software for digital video microscopy is developed based on the pose and depth estimation results with visualization of side views and 3D views of the microrobots, which is designed to assist the operator to conduct dexterous micro-manipulation.

The remainder of the paper is structured as follows. Firstly, the problem statement is presented while the methodology is illustrated in Section 2 for the depth estimation of micro-platforms as well as the pose and depth estimation of mobile microrobots. Section 3 describes the experimental setup, while the results analysis are presented. Finally, the conclusions are drawn and the future prospects are described in Section 4.

Figure 1: Illustration of the micro-platforms and the mobile microrobots used for the validation of the proposed algorithm in this paper, (top row images are CAD models and bottom row images are the indicative frames from the experimental datasets). Platform A and B are micro-platforms, and their poses will not change during experiments. Robot A and B represent a four-arm microrobot and a two-arm microrobot respectively, which may have various poses during micro-manipulation experiments.

2 Methodology

The proposed method aims to construct a precise and robust microscopic pose and depth estimation for optically transparent objects with complex shapes. Two types of objects are studied, namely micro-platforms (which are static objects) and mobile microrobots (which can be moved and have different out-of-plane poses during the operation). An illustration of the micro-platforms and the mobile microrobots is shown in Fig.1, which includes micro Platform A, micro Platform B, micro mobile Robot A and micro mobile Robot B.

2.1 2D Position Estimation

Given the microscopic video data, a bounding box is manually placed to identify the initial position of the micro-platform/microrobot of interest. Each image frame of the video is

cropped according to the bounding box. Gaussian filter is applied to remove the noises from the images. Subsequently, a binary segmentation of the micro-platform/microrobot is generated by applying a threshold operation to the video.

After pre-processing the image data, 2D positions are then computed from the centre of mass of this binary micro-platform/microrobot image after segmentation. The contrast of the images can be adjusted to make sure that the micro-platform can be segmented by the thresholding.

Suppose that $p(x, y)$ represents the pixel of each image frame, $C(x, y)$ represents the grey-scale level of that pixel. The centroid of the segmented area is $[x_c, y_c]$, which can be calculated as follows.

$$x_c = \frac{\sum_{x,y} C(x, y)x}{\sum_{x,y} C(x, y)}; y_c = \frac{\sum_{x,y} C(x, y)y}{\sum_{x,y} C(x, y)} \quad (1)$$

In this way, the 2D planar position of the micro-platforms/microrobots can be determined. A new video can be generated with reduced dimension, whose center point is aligned with the centroid of the microrobot. Therefore, the resulting images will only contain the features of interest, which speeds up the calculation as the data size is reduced without removing any necessary information. After determining the 2D position of the microrobot, depth estimation and pose estimation can be conducted.

2.2 Focus Measurement

For micro-platform, depth estimation is significant in visualizing the respective structure. Focus measurement information obtained from monocular microscope image sequences can be exploited to estimate the depth. The algorithm utilized to measure the focus level for every image pixel is usually referred to a focus measure (FM) operator. In order to estimate the translation along the z-axis (which is parallel to the optical axis) of the micro-

platform, the focus measurement information is utilized, which means that the mapping between the focus measurement for each image frame and the corresponding z-position of the micro-platforms/microrobots should be established. The full image is used for the focus measurement calculations.

The approximation of a single focus measurement metric cannot be used directly to reconstruct the z-displacement. Since two different depths can correspond to the same value of a specific focus measurement. Therefore, combining different metrics for depth regression is necessary. The metrics used for focus measurement are illustrated as follows.

For an image, entropy is related to the complexity contained of its content. Whenever the micro-platform is displaced from the focal plane increase or decrease, the entropy value changes accordingly. For entropy based measurement method, prior to estimating the translation along the z-axis, a calibration routine is required to establish the mapping between entropy and the displacement of the microrobot along the z-axis. Since a focused image is expected to have a higher information content, the entropy computed can be used as one of the focus measures.²⁸

Apart from the image entropy, the variance of the image Laplacian^{29,30} can be used, which represents the second derivative of an image. The Laplacian highlights regions of an image containing rapid intensity changes and are often used for edge detection. If an image has a high variance, then the image is likely to be in-focus, since it contains many edge-like features. On the other hand, if it is of very low variance, then the image is likely to be blurred, with less edges; hence, the object may be out of focus.

The focus measure based on an alternative definition of the Laplacian is proposed,³¹ which is defined based on the modified Laplacian. Suppose that $\Delta_m I(i, j)$ is the modified Laplacian of $p(x, y)$, computed as $\Delta_m I = |I * L_X| + |I * L_Y|$ ($*$ is the convolutional operator). The convolution masks used to compute the modified Laplacian are $L_X = [-1, 2, -1]$, and $L_Y = L_X^T$. Another popular focus measure based on the magnitude of image gradient is known as the Tenengrad focus measure.²⁸

The four focus measurement metrics (M1:entropy; M2: Laplacian Variance; M3: Modified Laplacian; M4: Tenengrad operator) are defined as follows.

$$M1 = -\sum_{k=1}^L P_k \log(P_k) \quad (2)$$

$$M2 = \sum_{x,y} (I(x,y) - \Delta I)^2 \quad (3)$$

$$M3 = \sum_{x,y} \Delta_m I(x,y) \quad (4)$$

$$M4 = \sum_{x,y} ((G_x(i,j))^2 + G_y(i,j)^2) \quad (5)$$

where P_k is the relative frequency of the k -th gray-scale level, L represents the total number of different gray-scale levels, I is the image Laplacian obtained by convolving $p(x,y)$ with the Sobel operator, ΔI is the mean value of the image Laplacian, G_x and G_y are the image gradients in x,y directions computed by convolving the given image $p(x,y)$ with the Sobel operator.

Focus measurement can be affected by the presence of image noises from different noise sources. Considering that image noise is unavoidable, while the robustness of statistics-based operators may not be stable under various image contrast and image saturation, combining multiple focus measure is essential for accurate depth estimation. Therefore, a statistic regression model based on the focus measures should be explored.

2.3 Gaussian Process Regression

Gaussian process regression (GPR) is a nonparametric Bayesian approach, which has many desirable properties and can be applied for depth estimation with digital video microscopy.³² It is data-efficient and is easy-to-obtain through simple parameterization. A natural Bayesian interpretation is used to allow uncertainty during prediction. Therefore, GPR is used for the regression of the depth model by considering the multiple focus measurement metrics.

$\mathbf{M}(p_t)(t = 1, 2, \dots, T)$ represents the focus measure results of the image p collected at time step t using metric $\mathbf{M} = [M_1, M2, \dots, M_n]^T (i = 1, 2, \dots, n)$, where n is the total number

of metrics, T is the total number of time steps. In this paper, $n = 4$ is used. $f(\cdot)$ is a targeted mapping function, which maps the focus measurement values to the depth value. Considering that the training data may be noisy, we assume that the regressed model contains additive independent Gaussian noise ϵ . Therefore, the regression model for the task with noise can be written as $z_t = f(\mathbf{M}(p_t)) + \epsilon$, where z_t represents the target depth values, $\epsilon \in \mathcal{N}(0, \sigma_n)$ follows a Gaussian distribution with zero mean and variance.

Consider the training set $D = \{(\mathbf{M}(p_t), z_t) | t = 1, 2, \dots, T\}$, A GPR model addresses the question of predicting the value of a scalar output z_t , given the new input p_t and the training data. Suppose that $\mathbf{M}(p^*)$ is known as the test observation, $f(\mathbf{M}(p^*))$ is the prediction result based on GPR. A Gaussian process prior, specified using a mean function and a covariance function is described in Equation (6).

$$f(\mathbf{M}(p)) \sim GP(\text{Mean}(\mathbf{M}(p)), K(\mathbf{M}(p), \mathbf{M}(p^*))) \quad (6)$$

where $\text{Mean}(\mathbf{M}(p)) = \mathbb{E}[f(\mathbf{M}(p))]$ represents the mean function, $K(\cdot)$ denotes the covariance function. A Radial Basis Function (RBF) is used as the kernel to generate the covariance function $K(\cdot)$. $K(\mathbf{M}(p), \mathbf{M}(p^*))$ can be calculated as follows.

$$K(\mathbf{M}(p), \mathbf{M}(p^*)) = \mathbb{E}[(f(\mathbf{M}(p)) - \text{Mean}(\mathbf{M}(p)))(f(\mathbf{M}(p^*)) - \text{Mean}(\mathbf{M}(p^*)))] \quad (7)$$

In the regression process, the main target is to make inferences about the relationship between inputs and targets, i.e. the conditional distribution of the targets given the inputs. As an infinite-dimensional multivariate Gaussian distribution, all the data in the training set are joint Gaussian distributed. The predictive distribution can be generated based on conditioning the joint Gaussian prior distribution on the observations.³³ After normalization to ensure prior mean to be zero, the prior of the distribution of joint $f(\mathbf{M}(p))$ and $f(\mathbf{M}(p^*))$ can be obtained. The predictive equations for the GPR model can be written as shown in

Equation(8).

$$f(X)|X^*, X, f(X^*) \sim \mathcal{N}(K(X^*, X)[K(X, X) + \mathbf{I}\sigma_n^2]^{-1}f(X), K(X^*, X^*) - K(X^*, X)[K(X, X) + \mathbf{I}\sigma_n^2]^{-1}K(X, X^*)) \quad (8)$$

where $X = M_i(p)$, $X^* = M_i(p^*)$.

In this way, the target value can be represented by the mean values $f(\mathbf{M}(p^*))$, and the variances in the covariance matrix of $f(\mathbf{M}(p^*))$ can indicate the uncertainties of the target values predicted by the model. More details of the calculation for GPR can be referred to³⁴.

2.4 Depth Estimation for Micro-Platform

For micro-platforms which are static, we assume that the regression function mapping from the images (or the focus measurement metrics) of the target micro-platforms to its relative depth levels follows a Gaussian process. Therefore, GPR can be used to regress the model. Since the GPR method is a data-efficient approach, a small dataset for each micro-platform is enough to regress the model. The model training process can also be known as a few-shot regression process, which maps the corresponding focus measurement metrics of the image sequence to the depth values.

During the depth calibration process, each micro-platform is placed on the sample holder while the focus plane is adjusted to obtain different images of the micro-platform with different relative depth values. The 2D position detection is used to obtain the region of interest of the micro-platform. Subsequently, the focus measurement is conducted to generate the data for depth regression with GPR model. The depth values (ground truth data) and the obtained focus measurement metrics are normalized between 0 and 1 before calibration. Fig. 2 (a) and (c) show the normalized focus measurement metrics for different normalized depth values for micro-platform A and micro-platform B respectively.

For offline analysis, more metrics should be used to improve the accuracy of the regression.

Figure 2: Overview of the focus measure results and the corresponding normalized depth level for micro-platforms depth estimation calibration. (a) The normalized focus measurement metrics of different normalized depth values for micro-platform A. (b) The visualization of the GPR results for micro-platform A using two features. (c) The normalized focus measurement metrics of different normalized depth values for micro-platform B. (d) The visualization of the GPR results for micro-platform B using two features.

However, for online application, two features are sufficient for the depth regression. The two features with higher variance are chosen to train the GPR model. The visualization of the GPR training results is shown in Fig. 2 (b) and (d) for micro-platform A and micro-platform B respectively.

Though the GPR method can be extended to depth estimation for microrobots and the underlying regression process can be explained, the accuracy is not high enough. Moreover, for mobile micro-robots, the calibration is required to be conducted with various out-of-plane

poses. A robust approach for mobile microrobots depth estimation with various out-of-plane poses should be developed.

2.5 Pose and Depth Estimation for Microrobots

Depth estimation and pose estimation are critical for monitoring and controlling micro-robotic systems, but it is very challenging to obtain accurate estimations from a single image, due to the loss of 3D information during the image capture process.³⁵ Compared to handcrafted features, which are implemented based on focus measurements, the feature maps learned by neural networks can describe a more generalized mapping between the image and the specific depth value or pose. Therefore, we investigate the use of deep learning based methods to construct the digital video microscopy for microrobots with various poses.

Deep Residual Network (ResNet) can achieve high classification accuracy by training with hundreds of layers of neurons and yield compelling performance,³⁶ and it can be used for depth and pose estimation. In a traditional CNN model, inputs are passed from nodes of each layer to the next, with adjacent layers connected by convolutional operators. ResNets extend CNNs by adding short-term memory to each layer of the network, which forces the new layer to learn something different from what the input has already learned. ResNet was chosen to be integrated in the digital video microscopy for microrobots with various poses.³⁶

The coordinate definition of microrobot A and microrobot B is shown in Fig. 3 (a) and (c) respectively. The microrobots can rotate along the X,Y and Z axes and the rotation angles along the axes are known as pitch angle, roll angle and yaw angle respectively. The rotation along Z axis can be easily observed, since Z axis is parallel to the optical axis. The rotation along X and Y axis is known as out-of-plane rotation, which is often difficult to be perceived by the operators.

The target of 6D pose estimation of microrobots includes planar position estimation (2D), depth estimation (1D), planar rotation angle estimation (1D) and out-of-plane pose estimation (2D). The 2D planar position estimation has been addressed in the previous

Figure 3: The coordinate definition and the illustration of out-of-plane poses of microrobot A and microrobot B. (a) The coordinate definition of microrobot A. (b) Various out-of-plane poses of microrobot A. (c) The coordinate definition of microrobot B. (d) Various out-of-plane poses of microrobot B.

section, so the main focus of the construction of the deep learning model is to realize depth level ($z_k, k = 1, 2, \dots, K$) classification, out-of-plane pose ($\alpha_i, \beta_j, i = 1, 2, \dots, I, j = 1, 2, \dots, J$) classification, as well as planar rotation angle ($\gamma_m, m = 1, 2, \dots, M$) classification, since all the data is in discrete form during data collection process. The basic model used for training is shown in Fig. 4 (a). The ranges of the rotation angles and the depth level are determined by the specific application scenario.

A representative dataset is built with ground-truth data, which spans across sufficient configuration spaces of the microrobot workspace. For each image collected, it has a corresponding label of $[\alpha_i, \beta_j, z_k]$, where α represents pitch angle, β represents roll angle, z is the depth value. The minimal displacement between the rotation of the X and Y axis is 10 degree, while the displacement along Z axis is 1 μm . $n = I * J * K$ number of unique classes for the

database are collected, where $I = [\max(\alpha) - \min(\alpha)]/10 + 1$, $J = [\max(\beta) - \min(\beta)]/10 + 1$, and $K = \max(z) - \min(z) + 1$.

To reduce the data required to be collected, the database for planar rotation angle estimation is self-generated for self-supervised learning.³⁷ A set of pre-designed geometric transformations (e.g. planar rotation for $\gamma_m (m = 1, 2, \dots, M)$ degree) are applied to images with various out-of-plane poses and depth values in the database, where $M = [\max(\gamma) - \min(\gamma)]/5$. To this end, the neural network is trained to recognize how many degrees each microrobot rotates in the Z axis by fixing all the parameters of Convolutional blocks and tuning the parameters of the global average pooling layer.

A Hierarchical Semi Supervised ResNet architecture is therefore constructed for the multi-task classification, which is implemented by two parallel neural networks with the same structure using the original database for the depth and out-of-plane pose estimation respectively, and the self-generated database is used for the planar rotation estimation, as shown in Fig. 4 (b). Semi Supervised Learning represents the combination of supervised learning and self-supervised learning in two hierarchies to reach the target of multi-task learning. In Hierarchy 1 of the neural network architecture, supervised learning is used for out-of-plane pose and depth estimation based on the labeled data. As for the planar rotation angle estimation, self-supervised learning is conducted based on the self-generated image data and labels.

Though deep learning based methods can achieve high classification accuracy for many applications, they need large dataset with diversified data for training. Moreover, the model trained on one microrobot will fail to estimate the pose and depth estimation of another microrobot. Therefore, domain adaption should be considered.

2.6 Hybrid Model with Few-Shot Calibration

For the microscopic pose and depth estimation of microrobots, an ideal training set would be one which encompasses all possible poses and depth values. However, it is almost impossible to

Figure 4: The ResNet model and the overall framework. (a) The details of the basic model. (b) The overview of the Hierarchical Semi-Self Supervised ResNet architecture which is targeted for multi-task learning of microrobots out-of-plane pose, depth and planar rotation estimation.

collect a complete database with all the targeted poses and depth values for every microrobot with specific shape. Moreover, the data collected for the database is often snapshot data in discrete form, which is non-sequential data. For micro-manipulation tasks using optical microrobots, the data is sequential in continuous form. This limits the accuracy for estimation when formulating the neural network model as a regression model. The classification model trained with data in discrete form may not work for continuous depth values estimation. To enhance the precision of regression, a hybrid model based on the combination of GPR and ResNet is proposed.

The advantages of the GPR-based method is that it incorporates prior knowledge and is explainable. However, it is pose-dependent and the root-mean-square-error (RMSE) of the regression is not high compared to the deep learning method, and the results depend significantly on the accuracy of the pose estimation results. For the microrobot at the same depth level, if the pose of the microrobot changes, the focus measurement values will change accordingly, which may lead to another depth value using the same regression model. As for the ResNet model, higher accuracy for depth estimation can be achieved for microrobots with various poses. However, the black-box effects of deep learning methods are difficult to be eliminated. Therefore, the combination of the two methods can complement each other and yield an explainable yet accurate pose estimation.

To further improve the accuracy of depth estimation when the microrobots are switching between different poses, both GPR and the ResNet model are used. The GPR-based method is pose-dependent, which means that different GPR regression models have to be trained for the same microrobot with different poses. Few-shot calibration means that for a microrobot with a specific pose, a series of images with various depth levels are collected to generate the GPR model.

Kalman filter is used to remove the noises for the real-time estimation results and fuse the information from the GPR and the ResNet models together. A process model is normally used to model the transformation of the process state, which can usually be represented as a linear stochastic difference equation. $x(t)$ represents the real depth value at timestep t , $v(t) = x(t+1) - x(t)$ represents the depth value increment. Therefore, the state-space function can be written as follows.

$$\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = A \begin{bmatrix} x(t-1) \\ v(t-1) \end{bmatrix} + \begin{bmatrix} \epsilon_g \\ \epsilon_d \end{bmatrix} \quad (9)$$

Figure 5: The overview of the pose and depth estimation for microrobots based on GPR-ResNet hybrid model with few-shot calibration.

where A is the transition matrix for state space, $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $X(t) = \begin{bmatrix} x(t) \\ v(t) \end{bmatrix}$ represents the state at time t , ϵ_g and ϵ_d are random variables representing the process noise.

Measurement model is defined to describe the relationship between the process state and the measurements. Suppose that the depth level estimated by the neural network is z_d , while the depth value regressed by the GPR model is z_g . Kalman filter is used to fuse the information together. The measurement equation can be expressed as follows.

$$\begin{bmatrix} z_d(t) \\ z_g(t+1) - z_g(t) \end{bmatrix} = H \begin{bmatrix} x(t) \\ x(t+1) - x(t) \end{bmatrix} + \begin{bmatrix} \zeta_g \\ \zeta_d \end{bmatrix} \quad (10)$$

where H is an identity matrix, ζ_g and ζ_d represent measurement noise to model the inaccuracy caused by the ResNet or the GPR model. In this way, the results estimated by the two models can be regarded as noisy sensor measurements of the digital video microscopy.

The overview of the pose and depth estimation for microrobots based on GPR-ResNet hybrid model with few-shot calibration is shown in Fig. 5. After pre-processing of the

microrobot with 2D planar position estimation, the focus measurement is conducted, while the out-of-plane pose and the depth level are determined via the ResNet model. Based on the out-of-plane pose estimation results, the specific GPR model is used to generate the depth increment value based on the focus measurement results. The depth level and the depth increment are fed into the Kalman filter, while the final depth value can be generated.

2.7 Domain Adaption

The major drawback of deep network are the long training time and the enormous computational expenses. **Therefore, to enable network adaptation, the residual network is trained with a big dataset, and transfer learning is used to fine-tune the network for pose estimation for other mobile microrobots with complex shapes.** To verify the effectiveness of domain adaption of the proposed method, further experiments are conducted.

To standardize the training, for the same microrobot, an array of microrobots printed at different poses is fabricated to generate the training dataset for one robot. The raw image data is used for the end-to-end training for pose estimation. **If the pose recognition is needed for another microrobot with different shapes, transfer learning can be applied to refine the network for the microrobot without the need to re-train the neural network model from scratch. Since the feature extraction processes are similar for pose estimation, the parameters for the convolutional layers are fixed, while the global average pooling layer and the fully connecting layer can be fine-tuned on other dataset with limited ground-truth data. In this way, the model trained based on the microrobot A can be adapted for microrobot B with relatively small refinements of the network. Similarly, the network model can be tuned to adapt to other microscopic systems with various point spread functions (PSF).**

3 Experiments

In this section, the microrobot fabrication process and the experimental setup are introduced. The depth estimation results for micro-robotic platforms are shown while the pose and depth estimation results for mobile microrobots are described.

3.1 Microrobot Fabrication

The Nanoscribe 3D-Printer (Nanoscribe, Germany) is used to fabricate the optical microrobots. Photoresist (Nanoscribe, IP-L 780) was selected as the material for microrobot fabrication. It is bio-compatible and dielectric, with a refractive index of 1.52, which is higher than the surrounding medium (the refractive index for water is 1.33).

Two-photon polymerization (2PP) method was used for the manufacturing process,³⁸ where the resolution is set to 100 *nm*, and realized by a 3D printing system (Nanoscribe GmbH, Germany). The micro-platforms and microrobots for experimental validation were printed on the glass substrate and placed in deionized (DI) water within a spacer. The DI water provides suitable medium for optical trapping and prevents the microrobots from attaching to the glass slide easily.

Sonication process is utilized to detach the microrobot from the glass substrate (coverlip). The parameters of the sonication process are tuned empirically with different levels of power and time duration. After the microrobot samples were dried, an optical microscope (Zeiss, UK) was used to visualize the microrobot array.

3.2 Experimental Setup

The proposed method can be implemented in all the systems which involve a microscope for optical microrobot monitoring and manipulation, while an Optical Tweezer (Elliot Scientific, UK) was used to verify the effectiveness of the proposed method. The micro-platforms and microrobots were imaged with a high-speed CCD camera (Basler AG, Germany) and an

oil immersion lens microscope (Nikon Ti) with 100x magnification. More details of the experimental setup is shown in Supplement 4 (Fig. S7).

The printed microrobots are kept fixed on the glass slide and placed on the piezo stage to generate the ground truth trajectories along the z-axis. The microrobots with various pose were translated using the piezo stage in discrete mode or continuous mode. The sinusoidal trajectory, triangular trajectory or random trajectory in z-axis can be generated with different moving frequency of the piezo stage, which are regarded as the continuous modes. For one specific poses, at least 1000 frames were collected for each sub-dataset. The live video of the microrobot was collected and processed offline. After the 2D position detection and estimation, the video data is cropped to the dimension of 256 X 256 pixels.

3.3 Results and Analysis

3.3.1 Depth Estimation for Micro-Platform

The regression model is effective when the focus plane is within a reliable range (smaller than the overall dimension of the micro-platform in the z axis). To verify that the proposed method can estimate the relative depth of the micro-platform compared to the focal plane, the translational stage is moved by following a pre-defined trajectory in z axis.

The depth of the 3D-printed micro-platforms is estimated by using the GPR model obtained from prior calibration. The Nanopositioner(Mad City Labs Inc.) has a working range of 100 μm . During the experiment, the microrobots are moving between the depth level of 46 μm and 52 μm . Three tests were conducted in total, using three microrobots fabricated at different time but have the same structure and printing parameters. Fig. 6 demonstrates the depth estimation results for the two micro-platforms respectively during Test 1. The comparisons between the experimental curve and the ground truth data of the other tests can be viewed in Supplement 2 (Fig. S2 and Fig. S3).

The root-mean-square-error (RMSE) is used to measure the accuracy of the regression. Three tests were conducted to verify the effectiveness of implementing the GPR model for

Figure 6: The validation of the GPR based depth regression in Test 1. (a) Experimental results for depth estimation using two selected focus measurement metrics for micro-platform A. (b) Depth estimation results using two selected focus measurement metrics for micro-platform B. (c) Experimental results of the depth estimation using four focus measurement metrics for micro-platform A. (d) Depth estimation results using four focus measurement metrics for micro-platform B.

depth estimation. RMSE results for the regression of micro-platform in all the tests are demonstrated in Table 1. It can be concluded that the regression accuracy can be significantly improved by using more features.

3.3.2 Pose and Depth Estimation for Microrobots

For 3D-printed microrobots of known geometries, a dataset can be generated consisting of a set of microscope images and the corresponding 3D positions and orientations of the microrobot. The labelled data is used to train the model, where 20% of the total images

Table 1: The RMSE results for the regression of micro-platforms.

2 features	Test 1	Test 2	Test 3	Mean $\pm$ Std
Micro-Platform A	0.739 μ m	0.935 μ m	0.809 μ m	0.828 $\pm$ 0.099
Micro-Platform B	0.649 μ m	0.555 μ m	0.645 μ m	0.616 $\pm$ 0.053
4 features	Test 1	Test 2	Test 3	Mean $\pm$ Std
Micro-Platform A	0.556 μ m	0.728 μ m	0.507 μ m	0.597 $\pm$ 0.116
Micro-Platform B	0.432 μ m	0.390 μ m	0.462 μ m	0.428 $\pm$ 0.036

was used for testing. During the model training process for the other 80% of data, 70% was used for training while 30% was used for validation to fine-tune the model. Detailed illustration of the 2D position detection process and the single image pre-processing for database construction can be found in Supplement 3 (Fig. S5).

To reduce the computational time, each image is resized to 50 $\times$ 50, while intensity of the image is normalized before feeding the images to the network. For both the training and testing data, the pre-processing of the image is conducted as follows.

$$p(x', y') = [p(x, y) - \frac{\sum_{x,y} p(x, y)}{2500}] / 128 \quad (11)$$

Data augmentation is used to avoid over-fitting during the training process. The batch size is set to be 100. The cross entropy loss is computed while AdamOptimizer is used as the optimizer. After training 30 epoches, the accuracy of pose estimation for the targeted classes can reach 100% for the testing database. The results of the pose estimation are shown in Fig. 7, including the top view obtained by the camera and the side view augmented by CAD model, which is generated with the pose estimation results. The same ResNet architecture can be used for the discrete depth level estimation for microrobots. The neural network can work simultaneously to generate the pose and depth estimation results.

The pose estimation accuracy for microrobot A is 99.93%. As for the depth estimation, the accuracy is 97.76% for microrobot A. The depth estimation results of microrobot A with various pose based on the DNN method can be viewed in Fig. 8 (a), the model of which is trained using the discrete dataset. The depth estimation results using the ResNet model for

Figure 7: Overview of the results for pose estimation. The top view of the image obtained by microscopic camera is fed to the neural network, the out-of-plane pose can be estimated while the additional side view can be visualized in the CAD model mode.

microrobot A and the comparisons with the ground-truth data is shown in Fig. 8 (c) for continuous depth value estimation.

As for the planar rotation angle estimation, the accuracy is 99.98%, which indicates that the self-supervised learning for Hierarchy 2 in the ResNet model is effective and the 6D pose estimation for the microrobot can be realized. The examples of planar rotation angle estimation are demonstrated in Fig. 9. Two indicative sets of feature maps derived from the first two convolutional layers and max pooling layers are shown in Supplement 3 (Fig. S6) to demonstrate the process of feature extractions by neural network.

3.3.3 Hybrid Model with Few-Shot Calibration and Domain Adaption

The model trained using the discrete data is adapted to the new dataset with many unseen data. The few-shot calibration is performed at first. Several examples for the focus mea-

Figure 8: Overview of the pose and depth estimation experimental results for the micro-robots. (a) The depth estimation results comparison between the ground-truth data and the ResNet model for the discrete dataset. (b) The few-shot calibration results using the GPR model. (c) Depth estimation results using the modified ResNet-18 for microrobot A and the comparisons with the ground-truth data. (d) Depth estimation results using the hybrid model (GPR-ResNet model) for microrobot A and the comparisons with the ground-truth data.

surement calibration of microrobot A are demonstrated in Supplement 1 (Fig. S1), while the experimental results of GPR model for depth estimation of microrobot A with different poses can be viewed in Supplement 2 (Fig. S4).

Fig. 8 (b) shows the results comparison between the ground-truth data and the depth estimation obtained based on GPR model for few-shot calibration of the small dataset for microrobot A. Combining the depth estimation results from both of the GPR and ResNet with Kalman filter, the depth estimation results using the hybrid model (GPR-ResNet model) for microrobot A and the comparisons with the ground-truth data is shown in Fig. 8 (d).

The quantitative results for the comparisons between the GPR model, ResNet model and the hybrid model are shown in Table 2. It can be clearly seen that with the hybrid model,

Figure 9: Overview of the planar rotation angle estimation via self-supervised learning based on the ResNet Model.

the RMSE for the depth estimation can be improved ($0.381\mu\text{m}$), compared with the results generated by the GPR model ($0.669\mu\text{m}$) and ResNet model ($0.447\mu\text{m}$).

Table 2: The experimental results of domain adaption for verification of the effectiveness of the proposed methods with comparative studies.

	GPR based Model	ResNet Model	Proposed Hybrid Model
Microrobot A	$0.669\mu\text{m}$	$0.447\mu\text{m}$	$0.381\mu\text{m}$
Microrobot B	$0.555\mu\text{m}$	$0.445\mu\text{m}$	$0.316\mu\text{m}$

The database for microrobot A is constructed for training the initial network. To accelerate the training speed, transfer learning is used to obtain the pose and depth estimation model for microrobot B. In order to test whether a small dataset is enough for the neural network to adapt to the pose estimation for another microrobot, only 20% of the data is used for training. With the small dataset, the pose estimation for microrobot B based on transfer learning is 99.93%, while the depth estimation accuracy is 95.47%, while the RMSE is $0.212\mu\text{m}$. The results demonstrate the adaptability of the proposed method for various types of microrobots.

Figure 10: The visualization of microrobot monitoring with various poses and depth values, including the top view obtained by microscopic camera and the side view with CAD model based depth level demonstration.

3.3.4 Visualization

With precise pose estimation and depth estimation, the monitoring of the microrobots with visualization can be realized to enable the operators to have a better perception of the microrobotics, as shown in Fig. 10. The blue line represents the focal plane. The side view of the microrobot is shown based on the pose estimation results, while the depth value is demonstrated by adjusting the relative position between the model and the blue line. The 3D model can overlaid to the microscopic images during operation.

An analytical software is used to provide more information to the users when constructing the digital microscopy. An example is shown in Fig.11, while the video for overall testing can be found in Supplementary Video S1. The evaluation metrics are calculated online and scaled down. The value of each metric is visualized in digital form, and the radius of the circle indicates the value. Therefore, the variation of the focus measurement can

Figure 11: An analytical software is used to provide more information to the users when constructing the digital microscopy. (a) The original figure of the microrobot under microscope. (b) The visualization of focus measurement metrics after proper scaling. (c) The 2D entropy map visualization. (d) The side view of the microrobot with depth estimation results (the unit is μm). (e) The top view of the microrobot with out-of-plane pose estimation results (the unit is degree). (f) The 3D view of the microrobot with planar rotation angle estimation results (the unit is degree).

be monitored (see Fig. 11(b)). The 2D entropy image is demonstrated to provide a more intuitive visualization of the focus measurement. Fig.11(d)-(f) are the visualization of the depth and pose estimation mentioned above.

4 Conclusions and Future Work

The aim of the proposed methodology is to develop deep learning method for the depth and pose estimation of micro-platforms and microrobots, and to enhance the perception of the operator with visualization of additional views. The depth can be regressed by utilizing multiple focus measurement information via GPR for the micro-platforms. A ground truth dataset is used to generate the model for depth regression. The effectiveness of the model is

tested on two micro-platforms for depth estimation, which is significant for the construction of digital microscope.

As for microrobot, the microrobot's depth value, out-of-plane and planar rotation angle are estimated with a deep learning method, using a Hierarchical Semi Supervised ResNet architecture. The 6D pose estimation can therefore be implemented. The model is trained based on a database with discrete labels. To reach the goal of continuous depth value estimation, a hybrid model based on GPR and ResNet is developed while Kalman filter is used to estimate the depth values smoothly. Transfer learning is tested by fine-tuning the model to realize the pose and depth estimation for another microrobot as validation. Visualization is used to enhance the operator's 3D perception of the microrobots during micro-manipulation.

Future work will include extending the method to pose and depth estimation of multiple microrobots simultaneously and demonstrate its usefulness in biomedical applications, such as indirect cell manipulation, which requires depth and pose information to enable multiple microrobots to work cooperatively.

Supporting Information

Supporting Information is available free of charge via the internet at <http://pubs.acs.org> or from the author.

Acknowledgements

The authors would like to acknowledge Dr M.Power and Dr F. Seichepine for the cleanroom training, would like to acknowledge Dr M.Grammatikopoulou and Dr A. Barbot for the optical tweezer system training.

References

- (1) Li, J.; de Ávila, B. E.-F.; Gao, W.; Zhang, L.; Wang, J. Micro/nanorobots for biomedicine: Delivery, surgery, sensing, and detoxification. *Science Robotics* **2017**, *2*.
- (2) Crocker, J. C.; Grier, D. G. Methods of digital video microscopy for colloidal studies. *Journal of colloid and interface science* **1996**, *179*, 298–310.
- (3) Jones, P. H.; Maragò, O. M.; Volpe, G. *Optical tweezers: Principles and applications*; Cambridge University Press, 2015.
- (4) Neuman, K. C.; Nagy, A. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature methods* **2008**, *5*, 491.
- (5) Waigh, T. A. Advances in the microrheology of complex fluids. *Reports on Progress in Physics* **2016**, *79*, 074601.
- (6) Sahoo, S. K.; Misra, R.; Parveen, S. *Nanomedicine in cancer*; Pan Stanford, 2017; pp 73–124.
- (7) Thompson, R. E.; Larson, D. R.; Webb, W. W. Precise nanometer localization analysis for individual fluorescent probes. *Biophysical journal* **2002**, *82*, 2775–2783.
- (8) Manley, S.; Gillette, J. M.; Lippincott-Schwartz, J. *Methods in enzymology*; Elsevier, 2010; Vol. 475; pp 109–120.
- (9) Hannel, M. D.; Abdulali, A.; O’Brien, M.; Grier, D. G. Machine-learning techniques for fast and accurate feature localization in holograms of colloidal particles. *Optics express* **2018**, *26*, 15221–15231.
- (10) Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J. A.; Van Ginneken, B.; Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis* **2017**, *42*, 60–88.

- (11) Newby, J. M.; Schaefer, A. M.; Lee, P. T.; Forest, M. G.; Lai, S. K. Convolutional neural networks automate detection for tracking of submicron-scale particles in 2D and 3D. *Proceedings of the National Academy of Sciences* **2018**, *115*, 9026–9031.
- (12) Helgadottir, S.; Argun, A.; Volpe, G. Digital video microscopy enhanced by deep learning. *Optica* **2019**, *6*, 506–513.
- (13) Stavroulakis, P.; Leach, R. K. Invited Review Article: Review of post-process optical form metrology for industrial-grade metal additive manufactured parts. *Review of Scientific instruments* **2016**, *87*, 041101.
- (14) Grammatikopoulou, M.; Zhang, L.; Yang, G.-Z. Depth estimation of optically transparent laser-driven microrobots. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2017; pp 2994–2999.
- (15) Marturi, N.; Tamadazte, B.; Dembélé, S.; Piat, N. Visual servoing-based depth-estimation technique for manipulation inside sem. *IEEE Transactions on Instrumentation and Measurement* **2016**, *65*, 1847–1855.
- (16) Cui, L.; Marchand, E.; Haliyo, S.; Régnier, S. Three-dimensional visual tracking and pose estimation in scanning electron microscopes. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2016; pp 5210–5215.
- (17) Marturi, N.; Tamadazte, B.; Dembélé, S.; Piat, N. Image-guided nanopositioning scheme for sem. *IEEE Transactions on Automation Science and Engineering* **2018**, *15*, 45–56.
- (18) Eichhorn, V.; Fatikow, S.; Wich, T.; Dahmen, C.; Sievers, T.; Andersen, K. N.; Carlson, K.; Bøggild, P. Depth-detection methods for microgripper based CNT manipulation in a scanning electron microscope. *Journal of Micro-Nano Mechatronics* **2008**, *4*, 27–36.

- (19) Kudryavtsev, A. V.; Dembélé, S.; Piat, N. Full 3d rotation estimation in scanning electron microscope. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2017; pp 1134–1139.
- (20) Bergeles, C.; Kratochvil, B. E.; Nelson, B. J. Visually servoing magnetic intraocular microdevices. *IEEE Transactions on Robotics* **2012**, *28*, 798–809.
- (21) Rivenson, Y.; Ceylan Koydemir, H.; Wang, H.; Wei, Z.; Ren, Z.; Gunaydin, H.; Zhang, Y.; Gorocs, Z.; Liang, K.; Tseng, D., et al. Deep learning enhanced mobile-phone microscopy. *Acs Photonics* **2018**, *5*, 2354–2364.
- (22) Wu, Y.; Ray, A.; Wei, Q.; Feizi, A.; Tong, X.; Chen, E.; Luo, Y.; Ozcan, A. Deep learning enables high-throughput analysis of particle-aggregation-based biosensors imaged using holography. *ACS Photonics* **2018**, *6*, 294–301.
- (23) Wu, Y.; Calis, A.; Luo, Y.; Chen, C.; Lutton, M.; Rivenson, Y.; Lin, X.; Koydemir, H. C.; Zhang, Y.; Wang, H., et al. Label-free bioaerosol sensing using mobile microscopy and deep learning. *ACS Photonics* **2018**, *5*, 4617–4627.
- (24) Grammatikopoulou, M.; Zhang, L.; Yang, G.-Z. Depth estimation of optically transparent microrobots using convolutional and recurrent neural networks. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2018; pp 4895–4900.
- (25) Mahendran, S.; Lu, M. Y.; Ali, H.; Vidal, R. Monocular object orientation estimation using Riemannian regression and classification networks. *arXiv preprint arXiv:1807.07226* **2018**,
- (26) Grammatikopoulou, M.; Yang, G.-Z. Three-dimensional pose estimation of optically transparent microrobots. *IEEE Robotics and Automation Letters* **2019**, *5*, 72–79.

- (27) Qu, Y.; Jing, L.; Shen, Y.; Qiu, M.; Soljacic, M. Migrating knowledge between physical scenarios based on artificial neural networks. *ACS Photonics* **2019**, *6*, 1168–1174.
- (28) Nelson, B. J. Autofocusing in computer microscopy: Selecting the optimal focus algorithm. *Microscopy Research and Technique* *65*, 139–149.
- (29) Fernandez-Valdivia, J. Diatom autofocusing in brightfield microscopy: A comparative study. Pattern Recognition, 2000. Proceedings. 15th International Conference on. 2000.
- (30) Pech-Pacheco, J. L.; Cristóbal, G.; Chamorro-Martinez, J.; Fernández-Valdivia, J. Diatom autofocusing in brightfield microscopy: a comparative study. Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. 2000; pp 314–317.
- (31) Nakagawa, Y. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence* *16*, 824–831.
- (32) Williams, C. K. I.; Rasmussen, C. E. Gaussian Processes for Regression. *Advances in Neural Information Processing Systems* **1996**, *8*.
- (33) Rasmussen, C. E. Gaussian processes in machine learning. Summer School on Machine Learning. 2003; pp 63–71.
- (34) Schulz, E.; Speekenbrink, M.; Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology* **2017**, *85*, 1–16.
- (35) Chang, J.; Wetzstein, G. Deep optics for monocular depth estimation and 3d object detection. Proceedings of the IEEE International Conference on Computer Vision. 2019; pp 10193–10202.
- (36) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

- (37) Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728* **2018**,
- (38) Kawata, S.; Sun, H.-B.; Tanaka, T.; Takada, K. Finer features for functional microdevices. *Nature* **2001**, *412*, 697.

Graphical TOC Entry