IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Fine-Grained Classification for Depth Estimation from Monocular Microscopy for Robotic Micromanipulation of Motile Cells

Han Yang[1,2], Yufei Jin[1,2], Aojun Jiang[3], Xinrui Wang[2],Xibu Wang[4], Xiaoling Yi[4], Yu Sun[3,*] and Zhuoran Zhang[1,2,*]

***Abstract*—Manipulation of motile cells is crucial for biological research and clinical applications. However, obtaining Z-axis visual feedback under monocular microscopy remains a challenge for robotic micromanipulation. Traditional depth-from-focus and depth-from-defocus methods fail to handle motile cells due to time-consuming focus search or inaccurate defocus modeling. This paper addresses these limitations by reformulating depth estimation as a fine-grained multi-class depth classification problem that exploits the shallow depth-of-field characteristic of microscopy. We propose a Fine-Grained Attention Fusion Module (FGAF-Module) that combines multi-scale grouped convolution for extracting subtle depth-related features with attention mechanisms to focus on discriminative regions in cell images. Additionally, channel-based feature augmentation methods, including CrossNorm and SelfNorm, enhance fine-grained feature discrimination while improving model generalization to handle morphological variations during cell movement. A weighted loss function further guides the model to distinguish between adjacent depth categories by penalizing errors proportionally to depth differences. For network training evaluation, the FGAF-module enhanced network achieved 83.52% top-1 classification accuracy and 96.88% top-3 classification accuracy while maintaining real-time performance at 90 frames per second. To demonstrate the capability of our approach in providing visual feedback for robotic manipulation of motile cells, the trained depth estimation model was integrated into a robotic sperm aspiration system. The model provided real-time visual depth feedback to guide 3D pipette localization during sperm aspiration procedures, achieving a 92% success rate for live motile sperm aspiration. These results validate the effectiveness of fine-grained classification for monocular depth estimation in micromanipulation applications.**

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

I. Introduction

Manuscript received: September, 2, 2025; Accepted November, 10, 2025.

This paper was recommended for publication by Editor Xinyu Liu upon evaluation of the Associate Editor and Reviewers' comments. 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, 62588301), in part by the Fundamental Research Funds for the Central Universities, and in part by Guangdong Basic and Applied Basic Research Foundation (2024A1515010160), all to Z. Zhang.

[1]The authors are with the Institute of Robotics and Intelligent Systems, Dalian University of Technology, Dalian, Liaoning, China.

[2]The authors are with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China.

[3] The authors are with the Robotics Institute, University of Toronto, Canada.

[4] The authors are with Department of Reproductive Medicine, The 3rd Affiliated Hospital of Shenzhen University, Shenzhen, China.

*Correspondence to Z. Zhang (zhangzhuoran@cuhk.edu.cn) and Y. Sun (yu.sun@utoronto.ca)

Digital Object Identifier (DOI): see top of this page.

©2026 IEEE

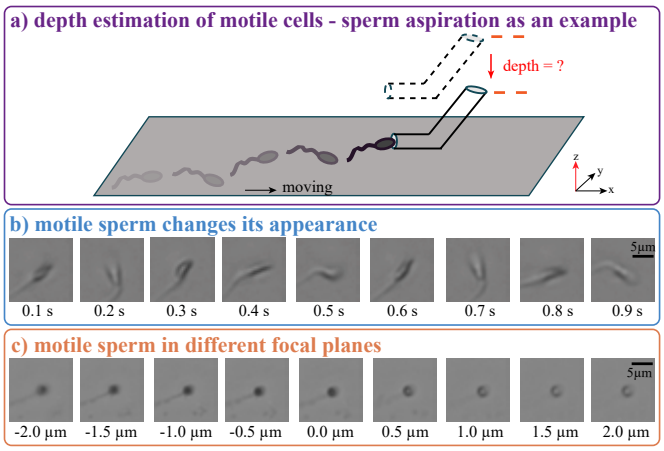

Fig. 1. **Challenges in monocular depth estimation for robotic micromanipulation of motile sperm**. a) Taking sperm aspiration as an example, precise depth estimation is critical as even minor misalignments lead to aspiration failure. b) First challenge: Rapid morphological changes during sperm motility create a many-to-one relationship between appearance and depth. c) Second challenge: Subtle differences in sperm projection features across adjacent focal planes complicate feature discrimination.

MANIPULATION of motile cells plays a critical role in biological and robotics research and clinical applications. In biological research, motile cell manipulation enhances understanding of cellular movement mechanisms and advances bio-hybrid micro-robot development through techniques including optogenetics [1], magnetic fields [2], electric fields [3], and fluidic fields [4]. In clinical applications such as in vitro fertilization (IVF) treatment, a motile sperm needs to be aspirated (see Fig. 1a) and immobilized before being injected into an egg cell [5]. These manipulation tasks are involved in over 70% of IVF treatments [6], which affect more than 50 million couples worldwide [7].

Motile cell manipulation is performed in three-dimensional space, and visual feedback of the 3D positions (XYZ) of the motile cell is a prerequisite for manipulation. Within the focal plane, visual feedback of XY positions can be obtained in real time through methods such as 2D image-based detection [8], [9] and segmentation algorithms [10], [11]. However, obtaining the Z-axis depth information under microscopes remains challenging. Conventional depth sensors such as LiDAR [12], [13] and RGB-D cameras [14], [15] cannot be integrated with microscopes. Alternative imaging modalities such as confocal microscopy require fluorescent staining of cells, which is not allowed in clinical practice [16]. Holographic microscopy not

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

only conflicts with standard clinical setups, but is also time-consuming to reconstruct the 3D distribution of targets (e.g., 4.1 seconds per image [17]), making it unsuitable for motile cell manipulation tasks.

Under standard clinical setups (monocular optical microscopy), existing depth estimation methods are divided into two categories: Depth from Focus (DFF) and Depth from Defocus (DFD). The nature of DFF is autofocus. It incrementally adjusts the focal plane to search for the most in-focus plane, then calculates relative depth based on the most focused plane position [18], [19]. The most in-focus plane is identified using either focus measure algorithms (such as Entropy or Tenengrad) [20], or neural networks [21] [22]. Although DFF is suitable for estimating depth information of immotile cells, its trial-and-error searching process proves too time-consuming for tracking rapidly moving cells. Differently, DFD methods construct a defocus model to map blur image features to depth values [23], [24]. While DFD avoids focal plane searching, its accuracy depends on the accuracy of the established defocus models.

For motile cells such as sperm, the intrinsic cell movement makes existing methods unsuitable for depth estimation, due to the following challenges.

- Motile sperm frequently change their appearance; hence, different sperm appearances may correspond to the same depth value (same focal plane). This creates a many-to-one relationship when mapping image features to depth values, complicating depth estimation model development (see Fig. 1b).
- Image features of sperm at different focal planes show subtle differences. Even for immotile sperm, it is difficult to visually distinguish differences between image features at adjacent depths (see Fig. 1c), calling for fine-grained feature extraction to discern the subtle differences for accurate depth estimation.

This paper tackles these challenges by developing a multi-class classification approach for motile cell depth estimation. Despite classification-based depth estimation has been explored in computer vision [25], this paper focuses on obtaining depth information of motile cells to guide micromanipulation. Our approach utilizes the shallow depth-of-field of microscopes and reformulates the depth estimation problem into a multi-depth classification task. A plug-and-play fine-grained attention fusion module is developed to address both challenges: the attention fusion module and feature augmentation methods handle morphological variations, while the fine-grained module and weighted loss function distinguish subtle differences between adjacent focal planes. Integrating the proposed module improves baseline network performance on multi-depth classification tasks and demonstrates effective visual guidance for manipulation tasks. Using sperm aspiration as an example application, the module was trained on sperm dataset and achieved top-1 and top-3 accuracies of 83.52% and 96.88%. To demonstrate the capability of the proposed method for providing visual feedback in real-world cell manipulation tasks, the trained network integrated with our module was incorporated into a robotic sperm aspiration system. The proposed method provided visual feedback of sperm depth in real time (90 FPS) and enabled robotic aspiration of live motile sperm with an aspiration success rate of 92%.

II. Problem Formulation

The core of depth estimation under monocular microscopes is establishing the mapping relationship between image features and depth values. As shown in (1), the mapping function f_θ captures the correspondence between image features I_d and depth information Z_d.

$$f_\theta : I_d \rightarrow Z_d \tag{1}$$

Traditional depth estimation methods (DFF and DFD) employ different technical approaches to establish this mapping. DFF identifies the focal plane where the target appears sharpest by sequentially moving the focal plane through the specimen. DFD analyzes blur characteristics at different depths to construct a defocus mapping model that relates blur patterns to depth values. However, DFF requires time-consuming focal adjustments, while DFD heavily depends on the accuracy of the established defocus model. These limitations motivate the exploration of alternative depth estimation approaches.

Microscopic imaging's inherently limited depth-of-field creates a natural mechanism for depth estimation. Only objects within a specific focal plane appear sharp, while those at other depths become progressively blurred. This optical characteristic allows continuous depth variations to be discretized into distinct focal planes.

$$depth-of-field = \frac{\lambda \cdot n}{NA^2} + \frac{n}{M \cdot NA} \cdot e \tag{2}$$

Equation (2) quantifies the depth-of-field in microscopy, where e represents the detector element size. For an objective lens with numerical aperture (NA) of 0.65 and magnification of 40×, the calculated depth-of-field approximates 1.0 μm. This narrow range defines the microscope's z-axis resolution, indicating the region where objects appear clearly focused. During sperm swimming, the rapid vertical movements of sperm cause portions to move outside the focal plane, resulting in partially blurred images.

To leverage this optical property for depth estimation, the continuous depth spectrum can be discretized into defined labels Z_d^*. This discretization approach transforms depth estimation into a classification problem where each depth label corresponds to a specific focal plane with distinct blur characteristics.

The depth estimation problem is thus reformulated as a multi-class classification task. As shown in (3), the mapping function f_θ becomes a classification function f_θ^* that associates image features I_d with discrete depth labels Z_d^*.

$$f_\theta^* : I_d \rightarrow Z_d^* \tag{3}$$

From a probabilistic perspective, this classification task estimates the conditional probability $P(Z_d^*|I_d)$ of image features belonging to each depth label:

$$P(Z_d^*|I_d) = \frac{P(I_d|Z_d^*)P(Z_d^*)}{P(I_d)} \tag{4}$$

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

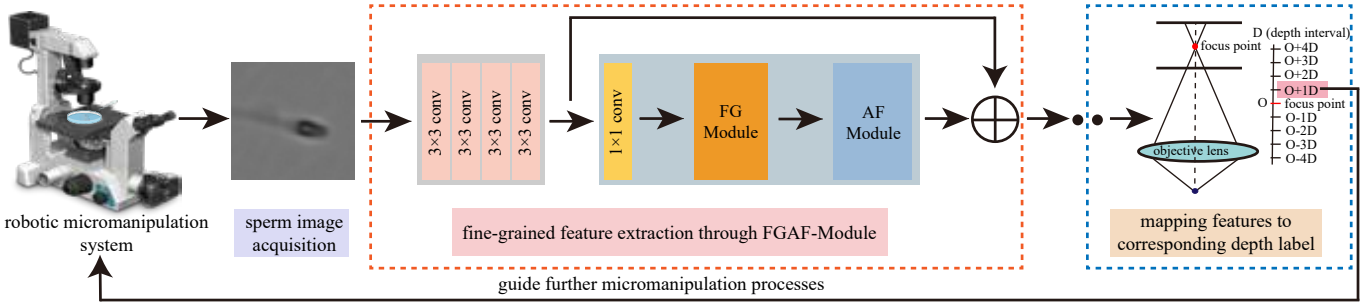

Fig. 2. **Depth estimation pipeline**. The proposed depth estimation pipeline for motile sperm consists of three main parts: 1) sperm image acquisition, 2) integration of the FGAF-Model for fine-grained feature extraction within a standard CNN, 3) mapping these features to depth labels to predict the sperm's depth for guiding micromanipulation. D represents the depth interval.

This Bayesian formulation maximizes the posterior probability of the true depth label. This problem is solved by developing neural networks that identify fine-grained features from sperm images and establish mappings between image features and depth labels, further formulating the problem as an optimization problem. The optimization objective determines the parameters θ^* that minimize the loss function L across the dataset $D = \{I_d, Z_d^*\}_{d=1}^N$:

$$\theta^* = \arg\min_{\theta} \sum_{d=1}^{N} L(f_\theta^*(I_d), Z_d^*). \tag{5}$$

This approach exploits the shallow depth-of-field in microscopy by reformulating continuous depth estimation as a discrete classification problem. The method aligns with the inherent properties of microscopic imaging while enabling precise depth estimation for micromanipulation applications.

III. Methodology

A. Overview

The proposed depth estimation pipeline for motile sperm consists of three primary components: sperm image acquisition, fine-grained feature extraction, and depth label classification (see Fig. 2). Initially, a robotic micromanipulation system captures images of motile sperm as input for the depth estimation network. Subsequently, the Fine-Grained Attention Fusion Module (FGAF-Module) extracts discriminative features from these images. Finally, the extracted features are mapped to corresponding depth labels, providing depth visual feedback that guides further manipulation tasks.

B. Fine-Grained Attention Fusion Module

Constructing an accurate mapping between sperm images and depth values requires overcoming two fundamental challenges. First, sperm at different focal planes exhibit minimal visual differences due to their small size and lack of texture. Second, the rapid morphological changes during sperm swimming create complex many-to-one relationships between morphological features and depth values. These challenges limit the effectiveness of conventional convolutional neural networks (CNNs) in the depth estimation tasks under microscope.

Standard CNNs, while effective for generic feature extraction, struggle with sperm images for several reasons. As the distance from the in-focus plane increases, the subtle differences between adjacent depth planes become increasingly difficult to discern. The lack of distinct textures and the rapid morphological changes of sperm further complicate feature extraction. Attempts to capture more abstract features by deepening the network lead to diminished sensitivity to crucial fine-grained details [26]. This indicates a need for specialized architectures that can effectively differentiate subtle depth-related features in sperm images.

To address these limitations, a FGAF-Module is proposed, which integrates a Fine-Grained Module (FG-Module) and an Attention Fusion Module (AF-Module) (see Fig. 3).

The FG-Module addresses the challenge of extracting subtle depth-related features from sperm images. Given the homogeneous texture and rapid morphological changes of sperm, conventional CNNs fail to capture the fine-grained differences between adjacent depth planes. The FG-Module overcomes this limitation by implementing group convolution, which divides feature maps along the channel dimension into multiple groups. Each group processes a specific range of channels independently, preventing interference between scales and enhancing the extraction of multi-scale fine-grained features. For the features of group i, a convolution operation extracts the original feature map F_{c_i}. To maintain computational efficiency, global pooling compresses the spatial information after group convolution, reducing parameter count while preserving essential information.

$$A_i = Sigmoid(W_2 \cdot \mathrm{GeLU}(W_1 \cdot \mathrm{GlobalPooling}(F_{c_i}))) \tag{6}$$

After multi-scale feature extraction and compression, activation functions GeLU and Sigmoid generate a feature weight vector A_i containing global semantic information (see (6)). To balance the importance of features across different scales, a Softmax function recalibrates the feature vector A_i. This normalization ensures that the sum of feature weights across all scales equals one, facilitating multi-scale feature integration. The recalibrated weights are then element-wise multiplied with the original feature map F_{c_i} to produce the weighted feature map M_i (see (7)):

$$M_i = \mathrm{Softmax}(A_i) \odot F_{c_i} \tag{7}$$

While the FG-Module effectively extracts multi-scale features, the key differences in sperm images across depth categories primarily manifest in the texture blurriness of critical

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

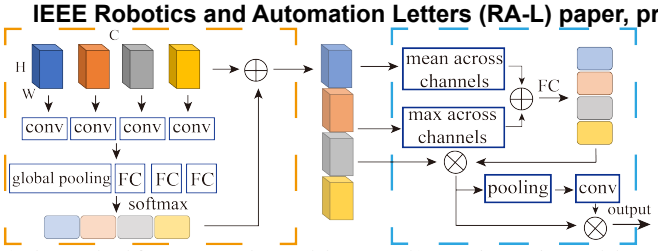

Fig. 3. **FGAF-Module**. The FGAF-Module consists of two components: the FG-Module and AF-Module. a) The FG-Module employs multi-scale grouped convolution to extract fine-grained features from sperm images. b) The AF-Module integrates attention mechanisms to enhance sensitivity to critical regions such as the sperm head.

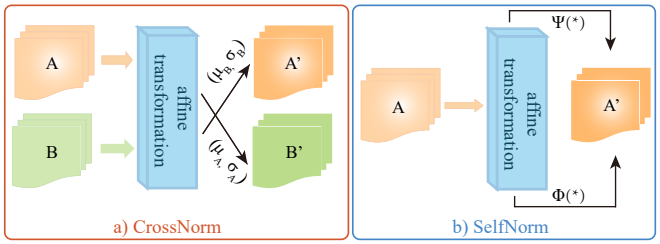

Fig. 4. **Channel-based Feature Augmentation**. a) CrossNorm exchanges mean and variance between feature maps of different channels, enabling the model to learn fine-grained features in sperm images. b) SelfNorm employs a learnable FCN and combines two attention mechanisms (Φ and Ψ) to adjust affine transformation parameters, bringing test data distribution closer to training data and improving model generalization.

regions such as the sperm head. Therefore, enhancing the model's sensitivity to these regions while suppressing less relevant information becomes essential for accurate depth estimation.

Conventional attention mechanisms [27], [28] primarily focus on feature channel importance, neglecting spatial details and feature interdependencies [29][30]. This limitation affects their ability to integrate local features with broader contextual information, which is crucial for capturing both the localized blur patterns in sperm heads and the overall morphological changes during swimming.

The AF-Module addresses this limitation by focusing on extracting key features within individual channels and adjusting spatial relationships to bridge local and global feature spaces. Taking the multi-scale fine-grained feature map M_i from the FG-Module as input, the AF-Module applies both average pooling and max pooling operations to each channel to capture typical and extreme features:

$$\text{Avg}_c = \text{mean}(M_i) \quad \text{Max}_c = \max(M_i), \quad \text{axis} = (H, W) \tag{8}$$

The pooled features are concatenated and processed through a fully connected layer to learn dependencies among features across channels, generating a channel importance weight matrix $F_{transformed}$ (see (9)). This matrix recalibrates the channel weights according to their assessed importance through element-wise multiplication with the original feature map M_i, resulting in the modified feature map M_i' (see (10)).

$$F_{\text{transformed}} = \text{FC}(\text{concat}(\text{Avg}_c, \text{Max}_c)) \tag{9}$$

$$M_i' = M_i \odot F_{\text{transformed}} \tag{10}$$

To bridge local and global feature spaces, M_i' undergoes spatial readjustment through pooling and convolution operations. This process reduces spatial dimensions to emphasize global features and establishes connections between localized fine-grained features and broader global features:

$$\text{Output} = \text{conv}(\text{Pooling}(M_i')) \tag{11}$$

Unlike existing attention mechanisms that primarily focus on channel relationships or spatial contexts independently, the FGAF-Module integrates both aspects. This integration enables the model to capture the subtle blur patterns in sperm heads while maintaining awareness of the overall morphological context, addressing both the fine-grained feature extraction challenge and the complex morphology-depth relationship in sperm images.

C. Channel-Based Feature Augmentation

To further enhance fine-grained features in sperm images, specialized feature enhancement methods are essential. Sperm images lack distinctive textures and become blurred away from the focal plane. Direct feature extraction causes models to learn noise rather than biological features, leading to overfitting.

Traditional data augmentation methods enhance image appearance rather than specific features. While these methods increase dataset size and improve generalization, they are insufficient for enhancing fine-grained features in sperm images given rapid morphological changes during swimming.

Two feature normalization methods, CrossNorm and SelfNorm, are integrated to address these limitations (Fig. 4). Unlike traditional methods, these approaches operate on feature layers by altering statistical properties of different channels, enhancing key features such as sperm head shape and blurring patterns.

CrossNorm is applied during training to enhance fine-grained features by exchanging statistical properties between channels of high-dimensional feature maps. Each channel's features $X \in \mathbb{R}^{H\times W\times 1}$ are normalized and transformed:

$$X = \alpha \frac{X - \mu_X}{\sigma_X} + \delta \tag{12}$$

By exchanging statistical properties between channels C and D:

$$C = \sigma_D \frac{C - \mu_C}{\sigma_C} + \mu_D, \quad D = \sigma_C \frac{D - \mu_D}{\sigma_D} + \mu_C \tag{13}$$

SelfNorm complements CrossNorm by operating during both training and testing to reduce distribution differences between training and test data. It employs attention mechanisms Φ and Ψ to adjust transformation parameters:

$$X = \Phi(\mu_X, \sigma_X)\sigma_X \frac{X - \mu_X}{\sigma_X} + \Psi(\mu_X, \sigma_X)\mu_X \tag{14}$$

CrossNorm expands feature distributions to improve discrimination between adjacent depth planes, while SelfNorm reduces distribution differences to handle morphological variations. This combined approach improves the model's ability to extract fine-grained features for accurate sperm depth estimation.

D. Weighted Loss Function

In fine-grained classification of sperm images at different depths, the traditional cross-entropy loss (CELoss) [31] func-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

tion proves inadequate for capturing subtle differences between adjacent depth categories. Sperm images collected based on depth gradients exhibit gradual changes in features from clear to blurry, with greater category differences corresponding to more pronounced differences in depth gradients and image blurriness. Models trained with CELoss tend to misclassify images during early training stages due to the minimal feature differences between adjacent depth categories.

To address this limitation and enhance classification accuracy, a weighted loss (WLoss) function is introduced, incorporating a penalty coefficient for incorrect predictions:

$$WLoss = -\frac{1}{N}\sum_{n=1}^{N} W_n \log(P_n, I), W_n = \left|{L_i}^2 - {L_n}^2\right| + 1 \quad (15)$$

where W_n represents the sample weight, L_i denotes the ground truth, L_n indicates the predicted value, and $P_{n,i}$ is the predicted probability output by Softmax. By assigning higher weights to samples with larger depth category differences, WLoss encourages the model to focus on learning the fine-grained features that distinguish between adjacent depth.

Unlike conventional loss functions that treat all classification errors equally, WLoss penalizes errors proportionally to the depth difference, reflecting the physical reality that larger depth differences correspond to more significant visual differences. This approach aligns the learning objective with the physical characteristics of the depth estimation problem, improving the model's ability to accurately classify sperm images across different depth gradients despite the subtle visual differences and rapid morphological changes.

IV. Experiments & Results

A. Dataset Preparation & Implementation Details

Data collection and micromanipulation in this work were performed using an inverted microscope (ECLIPSE Ti2, Nikon Inc.) with a 40× objective lens (NA = 0.65) for sample observation, equipped with a CMOS camera (acA1920-40u, Basler Inc.) to provide visual feedback (Fig. 5). The platform incorporated a motorized XY-stage (ProScan, Prior Scientific Inc.) with a 75 mm travel range, a 3-DOF (X-Y-Z translation) motorized micromanipulator (uMp-285, Sensapex Inc.) for precise micropipette localization, and an oil pump (CellTram Vario, Eppendorf Ltd.) for sperm aspiration. Computational tasks were executed on a computer with a GPU 3070Ti, handling both the user interface and neural network operations.

The networks were trained on custom collected dataset comprising 2,088 sperm images (232 image stacks from different sperm, with each stack containing the same sperm's 9 images captured at intervals of 0.5 μm from -2 μm to +2 μm, totaling 2,088 images with a resolution of 96×96 pixels). During training, a batch size of 64 was employed with an SGD optimizer, incorporating a weight decay of 1e-4, momentum of 0.9, and an initial learning rate of 0.01.

All samples were collected from the Department of Reproductive Medicine, The 3rd Affiliated Hospital of Shenzhen University, Shenzhen, China, with prior approval obtained from the institutional review board/ethics committee of the hospital and informed consent obtained from all participants.

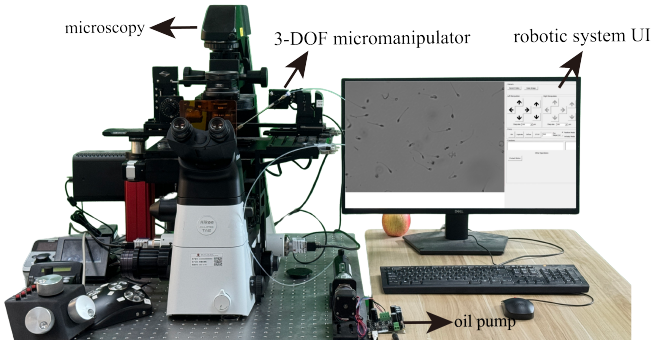

Fig. 5. **System setup**. The system includes a microscope, a 3-DOF micromanipulator, an oil pump, and a computer. Micromanipulation is performed through the system's user interface.

B. Network Performance

The developed FGAF-Module demonstrates versatile integration capability with various backbone architectures. To evaluate its effectiveness, the module was integrated with multiple network structures and subjected to quantitative performance analysis. As shown in Table I, the integration consistently enhanced both top-1 accuracy and top-3 accuracy across all tested networks.

ResNet-50 [32] with the FGAF-Module achieved the highest performance with 83.52% top-1 accuracy and 96.88% top-3 accuracy. This represents a 13.92% accuracy improvement over the baseline ResNet-50, demonstrating the effectiveness of the fine-grained feature extraction strategy for sperm depth discrimination. The marked enhancement confirms that the module improves the network's ability to discern subtle feature differences at various depth planes.

MobileNetv2 [33] augmented with the FGAF-Module also demonstrated notable performance gains. Despite its compact structure with only 4.6M parameters—significantly fewer than ResNet architectures—it achieved 73.29% top-1 accuracy, confirming that lightweight networks effectively leverage the FGAF-Module for improved performance. The inference time of MobileNetv2 with FGAF remained at just 3 ms compared to ResNet-50's 13 ms, highlighting the practical utility of this configuration for real-time applications without compromising accuracy.

The ShuffleNet [34] variants similarly benefited from FGAF integration. With the ShuffleNet_v2_×1.0 version, approximately 2% improvement in top-1 accuracy and over 1% increase in top-3 accuracy was observed. The inference time remained unchanged, indicating that the FGAF-Module maintains computational efficiency while enhancing feature extraction capabilities. The more powerful ShuffleNet_v2_×2.0 with FGAF achieved 80.19% top-1 accuracy, demonstrating that even efficient architectures can approach the performance of larger networks when equipped with the developed modules.

Furthermore, to evaluate the physical accuracy of the developed FGAF-Module integrated with various backbone architectures for providing visual feedback to robotic systems, the average depth estimation error between predicted and ground truth depth values was assessed. Although the developed model performs discrete depth classification, the depth estimation error metric demonstrates the real-world

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
NETWORK PERFORMANCE

Model Name	Top-1 Acc. (%)	Top-3 Acc. (%)	Params (M)	Time (ms)	Depth Estimation Error	
					Mean±Std (μm)	Var (μm^2)
ResNet-18	67.61	92.67	11.7	8	0.44±0.18	0.032
ResNet-18 with FGAF	72.28	93.36	13.5	9	0.26±0.11	0.012
ResNet-34	68.26	93.92	21.8	9	0.39±0.16	0.026
ResNet-34 with FGAF	81.47	94.44	23.2	11	0.15±0.07	0.005
ResNet-50	69.59	95.32	25.6	11	0.32±0.14	0.020
ResNet-50 with FGAF	**83.52**	**96.88**	28.3	13	**0.14±0.06**	**0.004**
MobileNetv2	68.84	91.74	3.5	2	0.37±0.15	0.023
MobileNetv2 with FGAF	**73.29**	**92.65**	4.6	3	**0.21±0.09**	**0.008**
GoogleNet	64.08	92.42	13.0	13	0.54±0.22	0.048
GoogleNet with FGAF	**67.92**	**93.51**	13.9	16	**0.41±0.17**	**0.029**
Shufflenet_v2_×0_5	72.25	91.61	1.4	3	0.26±0.11	0.012
Shufflenet_v2_×0_5 with FGAF	75.47	92.87	1.7	4	0.19±0.08	0.006
Shufflenet_v2_×1_0	75.84	91.44	2.3	5	0.19±0.08	0.006
Shufflenet_v2_×1_0 with FGAF	78.24	93.53	3.1	5	0.18±0.08	0.006
Shufflenet_v2_×1_5	76.79	92.95	3.5	6	0.19±0.08	0.006
Shufflenet_v2_×1_5 with FGAF	79.23	93.25	4.2	7	0.17±0.07	0.005
Shufflenet_v2_×2_0	78.98	93.01	7.4	8	0.18±0.08	0.006
Shufflenet_v2_×2_0 with FGAF	**80.19**	**94.26**	8.1	9	**0.16±0.07**	**0.005**

TABLE II
ABLATION STUDY FOR FGAF-MODULE & FEATURE AUGMENTATION METHOD

Model Name	FGAF-Module		Feature Augmentation		Loss		Top-1 Acc. (%)	Top-3 Acc. (%)
	FG-Module	AF-Module	CrossNorm	SelfNorm	CELoss	WLoss		
ResNet-34	✓	✓	✓	✓		✓	**81.47**	**94.44**
ResNet-34	✓		✓	✓		✓	74.28	93.36
ResNet-34		✓	✓	✓		✓	78.81	94.32
ResNet-34	✓	✓	✓	✓	✓		79.26	94.21
ResNet-34	✓	✓	✓			✓	78.37	94.02
ResNet-34	✓	✓		✓		✓	76.25	93.53

depth prediction accuracy by calculating the average difference between predicted and ground truth depth values across test samples when providing Z-axis visual feedback. Statistically, the results in Table I presented depth estimation error across all network architectures. ResNet-50 with FGAF achieved the lowest average depth estimation error of 0.14 μm with minimal variance of 0.004 μm^2, representing a 56.3% reduction compared to the baseline ResNet-50 (0.32 μm). Similarly, MobileNetv2 with FGAF demonstrated a 43.2% improvement in depth estimation error reduction (from 0.37 μm to 0.21 μm) while maintaining computational efficiency. The ShuffleNet variants showed consistent depth estimation error improvements, with ShuffleNet_v2_×2.0 with FGAF achieving 0.16 μm average error and reduced variance. This work employed a physical depth interval of 0.5 μm along the Z-axis for dataset collection and model training. The achieved average depth estimation error of 0.14 μm, lower than the depth interval, resulted from averaging all depth estimation errors across test samples, similar to how the top-3 classification accuracy of 96.88% implies concentrated predictions around correct depths, as evidenced by the minimal variance values. While feature differences between adjacent depth planes are subtle but distinguishable, the depth interval selection represents a trade-off: smaller intervals would result in more subtle inter-depth image features, increased classification difficulty, more depth categories, and consequently larger prediction errors with reduced Z-axis visual feedback precision; whereas larger intervals would enhance feature discriminability and classification accuracy but compromise the overall precision of Z-axis visual feedback for robotic applications.

These results confirm that the FGAF-Module consistently enhances performance across network architectures of varying complexity, particularly for the fine-grained multi-class classification required in sperm Z-axis localization. The module's ability to improve both classification accuracy and depth prediction precision with minimal computational overhead makes it suitable for robotic applications.

C. Ablation Study

An ablation study was performed to evaluate the contribution of individual components within the FGAF-Module, feature augmentation methods, and loss functions to sperm depth prediction accuracy. Table II presents results using backbone ResNet-34.

The complete model with both FG-Module (7) and AF-Module (11), along with CrossNorm (13), SelfNorm (14), and WLoss (15), achieved 81.47% top-1 accuracy and 94.44% top-3 accuracy. Using only FG-Module with CrossNorm and SelfNorm resulted in 74.28% top-1 accuracy, while only AF-Module achieved 78.81%. The larger performance drop when removing AF-Module highlights the critical role of attention mechanisms for focusing on key regions in sperm images.

Replacing WLoss with standard CELoss decreased performance to 79.26% top-1 accuracy, confirming WLoss's advantage in guiding the model to focus on relative depth differences between categories.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE III
SPERM ASPIRATION EXPERIMENTAL RESULTS.

Cell Type	Method	Aspiration Times	Success Times	Failure Times	Success Rate
Immotile	DFF	50	48	2	96%
	Ours	50	49	1	98%
Motile	DFF	NA	NA	NA	NA
	Ours	50	46	4	92%

Feature augmentation methods showed significant contributions. CrossNorm alone achieved 78.37% top-1 accuracy, while SelfNorm alone yielded 76.25%. CrossNorm provides greater benefit through enhanced fine-grained feature discrimination via statistical property exchange between channels. However, combining both methods achieved optimal results (81.47%), demonstrating their complementary nature in addressing sperm depth estimation challenges.

These results demonstrate that each component addresses specific challenges in sperm depth estimation, with the complete version creating a robust framework for accurate depth estimation.

D. Cell Aspiration Experiments

To demonstrate the capability of the developed modules in providing visual feedback for robotic cell manipulation tasks, the trained depth estimation model (ResNet-34 with the FGAF-Module) was deployed into a robotic sperm aspiration system. Aspiration experiments were first conducted using immotile sperm, followed by aspiration of motile sperm as a more challenging test case.

Each aspiration experiment followed a standardized protocol where the motorized XY-stage maintained a consistent distance of 100 pixels between the target sperm and micropipette tip. A pre-trained UNeXt model [35] provided XY-plane position information through sperm head segmentation, while the depth network and DFF methods predicted the Z-axis position. Based on this three-dimensional position feedback, the micropipette was automatically positioned to the corresponding depth. Throughout the aspiration process, a constant pump flow rate of 1.2 nL/s was maintained. Successful aspiration was defined as retaining the sperm within the pipette's field of view after being aspirated from outside the pipette into the viewing area.

1) Immotile Sperm Aspiration: Following the standardized aspiration protocol, both DFF methods and the proposed network guided pipette movement to target depths for aspiration. Both methods achieved comparable success rates in 50 trials: 96% for DFF and 98% for the proposed method (see Table III). For traditional DFF methods providing visual depth feedback, the gray level gradient algorithm was selected for focus search. Comparative evaluation of different focus measurement algorithms [36] revealed that the gray level gradient, Laplacian, and Tenengrad algorithms successfully reached local maxima at the sharpest focal plane (z = 0 μm), while the Entropy and Brenner gradient methods exhibited errors (see Fig. 6a). The subtle blur differences between adjacent depth intervals cause focus measurement algorithms to misidentify near-focus depths as the optimal focal plane, leading to prediction errors when providing depth feedback. Failure cases with both

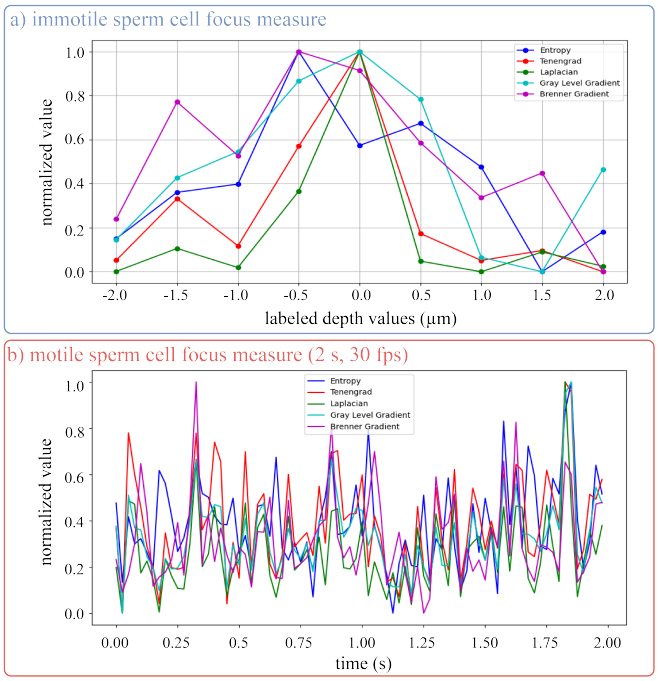

Fig. 6. a) For immotile sperm, five focus-measure algorithms (Entropy, Tenengrad, Laplacian, Gray level gradient, and Brenner gradient) were compared. b) For motile sperm, however, all methods are ineffective due to the sperm's inherent movement.

DFF methods and the developed depth estimation approach occurred when slight discrepancies between predicted and actual Z-axis positions caused the sperm to pass beneath the micropipette tip during aspiration attempts.

2) Motile Sperm Aspiration: For motile sperm, conventional DFF methods proved ineffective due to continuous morphological changes and shifting focus positions. As illustrated in Fig. 6b, the focus measurement algorithms failed to identify reliable peak focus values, rendering them unable to determine accurate Z-axis positions for motile sperm. Using the developed depth estimation method for aspiration experiments, a 92% success rate was achieved across 50 trials (see Table III).

The failure cases (4 out of 50) resulted from rapid depth changes of the swimming sperm, where morphological variations led to prediction errors. Inaccurate depth feedback caused misalignment between the micropipette tip and sperm, resulting in longer distances for the sperm to travel from outside the pipette tip to within the field of view. This led to the sperm swimming beneath the micropipette tip or, as the sperm approached the pipette tip where fluid velocity increased, the sperm entered the pipette with high velocity and quickly disappeared from the field of view. In rare instances, sperm suddenly changed their direction when approaching the micropipette tip, escaping from the aspiration range. For such failure cases, it is recommended to re-target the escaped sperm by timing the aspiration when the sperm is swimming toward the pipette tip, which reduces the escape probability.

These experimental results demonstrate that although conventional methods achieved comparable aspiration success rates with immotile sperm, the sperm movement made them incapable of providing depth information for motile sperm. they could not handle motile sperm. In contrast, the proposed depth

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

estimation network provides reliable Z-axis visual feedback for robotic micromanipulation of both immotile and motile sperm.

V. Conclusion

This work presents an approach for monocular depth estimation of motile cells in microscopy. The depth estimation problem was reformulated as a multi-class fine-grained classification task. A FGAF-Module was proposed to address the challenges of subtle visual differences across focal planes and complex morphology-depth relationships during cell movement. Fine-grained feature extraction and attention mechanisms were utilized to distinguish subtle feature differences. Experimental results demonstrate performance with 83.52% top-1 accuracy and 96.88% top-3 accuracy on sperm depth estimation while maintaining real-time processing at 90 FPS. Practical validation through robotic sperm aspiration achieves 98% success rates for immotile sperm and 92% for motile sperm, significantly outperforming conventional methods that fail entirely with motile cells. These results validate the effectiveness of fine-grained classification for providing accurate Z-axis visual feedback in robotic micromanipulation systems.

References

[1] X. Dong, S. Kheiri, Y. Lu, Z. Xu, M. Zhen, and X. Liu, "Toward a living soft microrobot through optogenetic locomotion control of caenorhabditis elegans," *Science Robotics*, vol. 6, no. 55, p. eabe3950, 2021.

[2] Q. Cao, Q. Fan, Q. Chen, C. Liu, X. Han, and L. Li, "Recent advances in manipulation of micro-and nano-objects with magnetic fields at small scales," *Materials Horizons*, vol. 7, no. 3, pp. 638–666, 2020.

[3] 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, no. 1, pp. 181–203, 2019.

[4] M. J. Doyle, J. V. A. Marques, I. Vandermeulen, C. Parrott, Y. Gu, X. Xu, A. Kolling, and R. Groß, "Modular fluidic propulsion robots," *IEEE transactions on robotics*, vol. 37, no. 2, pp. 532–549, 2020.

[5] L. Shingshetty, N. J. Cameron, D. J. Mclernon, and S. Bhattacharya, "Predictors of success after in vitro fertilization," *Fertility and Sterility*, vol. 121, no. 5, pp. 742–751, 2024.

[6] R. Maggiulli, D. Cimadomo, G. Fabozzi, L. Papini, L. Dovere, F. M. Ubaldi, and L. Rienzi, "The effect of icsi-related procedural timings and operators on the outcome," *Human Reproduction*, vol. 35, no. 1, pp. 32–43, 2020.

[7] N. E. Rusanova, " infertility and fertility: demographic problems of assisted reproduction." *Population & Economics*, vol. 8, no. 1, 2024.

[8] P. Hidayatullah, X. Wang, T. Yamasaki, T. L. Mengko, R. Munir, A. Barlian, E. Sukmawati, and S. Supraptono, "Deepsperm: A robust and real-time bull sperm-cell detection in densely populated semen videos," *Computer Methods and Programs in Biomedicine*, vol. 209, p. 106302, 2021.

[9] S. Zou, C. Li, H. Sun, P. Xu, J. Zhang, P. Ma, Y. Yao, X. Huang, and M. Grzegorzek, "Tod-cnn: An effective convolutional neural network for tiny object detection in sperm videos," *Computers in Biology and Medicine*, vol. 146, p. 105543, 2022.

[10] W. Dai, Z. Wu, R. Liu, T. Wu, M. Wang, J. Zhou, Z. Zhang, and J. Liu, "Automated non-invasive analysis of motile sperms using sperm feature-correlated network," *IEEE Transactions on Automation Science and Engineering*, 2024.

[11] W. Chen, H. Song, G. Shan, C. Dai, H. Liu, A. Jiang, C. Sun, C. Ru, C. Librach, Z. Zhang, *et al.*, "Automated parts segmentation of sperm via a contrastive learning-based part matching network," *IEEE Transactions on Automation Science and Engineering*, 2025.

[12] S. Shao, Z. Pei, W. Chen, Q. Liu, H. Yue, and Z. Li, "Sparse pseudo-lidar depth assisted monocular depth estimation," *IEEE Transactions on Intelligent Vehicles*, 2023.

[13] J. Choe, K. Joo, T. Imtiaz, and I. S. Kweon, "Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4672–4679, 2021.

[14] L. Papa, P. Russo, and I. Amerini, "D4d: An rgbd diffusion model to boost monocular depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[15] C. Liu, J. Gu, K. Kim, S. G. Narasimhan, and J. Kautz, "Neural rgb (r) d sensing: Depth and uncertainty from a video camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 986–10 995.

[16] J. Jonkman, C. M. Brown, G. D. Wright, K. I. Anderson, and A. J. North, "Tutorial: guidance for quantitative confocal microscopy," *Nature protocols*, vol. 15, no. 5, pp. 1585–1611, 2020.

[17] H. Wang, K. Bai, J. Chen, Q. Shi, T. Sun, J. Cui, Q. Huang, and T. Fukuda, "Digital holography based three-dimensional multi-target locating for automated cell micromanipulation," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 332–342, 2022.

[18] Z. Wang, C. Feng, W. T. Ang, S. Y. M. Tan, and W. T. Latt, "Autofocusing and polar body detection in automated cell manipulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 5, pp. 1099–1105, 2016.

[19] W. Liu and X. Wu, "Semi-global depth from focus," in *2015 3rd IAPR asian conference on pattern recognition (ACPR)*. IEEE, 2015, pp. 624–629.

[20] Y. Sun, S. Duthaler, and B. J. Nelson, "Autofocusing in computer microscopy: selecting the optimal focus algorithm," *Microscopy research and technique*, vol. 65, no. 3, pp. 139–149, 2004.

[21] J. Song, J. Wu, and K. Yu, "Learning-based auto-focus and 3d pose identification of moving micro-and nanowires in fluid suspensions," *IEEE Transactions on Automation Science and Engineering*, 2024.

[22] Z. Wang, H. Gong, K. Li, B. Yang, Y. Du, Y. Liu, X. Zhao, and M. Sun, "Simultaneous depth estimation and localization for cell manipulation based on deep learning," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 10 432–10 438.

[23] K. Taute, S. Gude, S. Tans, and T. Shimizu, "High-throughput 3d tracking of bacteria on a standard phase contrast microscope," *Nature communications*, vol. 6, no. 1, pp. 1–9, 2015.

[24] A. Zhang and J. Sun, "Joint depth and defocus estimation from a single image using physical consistency," *IEEE Transactions on Image Processing*, vol. 30, pp. 3419–3433, 2021.

[25] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[26] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artificial intelligence review*, vol. 53, pp. 5455–5516, 2020.

[27] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *ICCV*, 2021, pp. 783–792.

[28] S. Li, Q. Yan, and P. Liu, "An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism," *IEEE Transactions on Image Processing*, vol. 29, pp. 8467–8475, 2020.

[29] X. Yang, "An overview of the attention mechanisms in computer vision," in *Journal of physics: Conference series*, vol. 1693, no. 1. IOP Publishing, 2020, p. 012173.

[30] D. Soydaner, "Attention mechanism in neural networks: where it comes and where it goes," *Neural Computing and Applications*, vol. 34, no. 16, pp. 13 371–13 385, 2022.

[31] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International conference on Machine learning*. pmlr, 2023, pp. 23 803–23 828.

[32] H. et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

[34] Z. et al., "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018, pp. 6848–6856.

[35] V. et al., "Unext: Mlp-based rapid medical image segmentation network," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2022, pp. 23–33.

[36] S. et al., "Macro-to-micro positioning and auto focusing for fully automated single cell microinjection," *Microsystem Technologies*, vol. 27, pp. 11–21, 2021.