AI-enabled live-dead cell viability classification and motion forecasting

Anzhe Cheng[1], Chenzhong Yin[1], Michael A.S. Lamba[2], Mathieu Sertorio[2], Maldonado DeJesus[3], Jorge Alexis[3], Alexandre R. Sathler[4], Yu Chang[6], Catalin Chiritescu[4], Catherine A. Best[3], Dan Ionascu[2], Nicholas Kotov[5], Shahin Nazarian[1], and Paul Bogdan[1,*]

[1]Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90007, USA.
[2]Department of Radiation Oncology, College of Medicine, University of Cincinnati, Cincinnati, OH 45267, USA.
[3]College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
[4]Phi Optics, Inc., Champaign, IL 61820, USA.
[5]Department of Chemical Engineering, University of Michigan, Ann Arbor, MI 48109, United States
[6]Department of Statistics, The University of British Columbia, Vancouver, BC V6T 1Z4, Canada
[*]Correspondence and requests for materials should be addressed to P.B. (email: pbogdan@usc.edu)

ABSTRACT

Distinguishing live from dead cells is critical for studying cellular dynamics, monitoring of cell metabolism, and tissue heterogeneity studies. The cost of failure or uncertainty for live-dead image analysis can be particularly high for detecting disease emergence and evaluating medical therapies. Here, we present a novel deep learning framework that integrates a self-attention UNet for segmentation and a transformer network for dynamic tracking of cell movements. Our proposed model achieves state-of-the-art performance, with a high intersection-over-union (IoU) score of **96%** and an area-under-curve (AUC) score of **99%** for cell segmentation and over 65% IoU of full image cell motion forecasting, highlighting its ability to predict cell dynamics accurately. The self-attention mechanism significantly enhances the model's ability to differentiate live and dead cells, even in densely packed or morphologically diverse environments. Additionally, the transformer network effectively captures temporal dependencies, enabling precise predictions of future cell movements. This integrated framework demonstrates robust performance across diverse datasets, consistently outperforming existing methods. By offering high-accuracy segmentation and predictive modeling, our approach provides a transformative tool for advancing cellular analysis in research and clinical applications, including cancer diagnostics, drug development, and regenerative medicine.

1 Introduction

Accurate cell classification, particularly distinguishing live from dead ones, is essential in applications such as medical diagnostics, drug development, and tissue regeneration studies.[1–4] In cancer diagnostics, precise segmentation identifies tumor boundaries and assesses tumor growth dynamics, directly influencing treatment planning and prognosis.[5,6] This process allows radiologists to target cancerous regions accurately during radiation therapy.[7] In drug development, segmentation techniques evaluate cell viability and proliferation rates, offering insights into the efficacy and cytotoxicity of new compounds.[8–10] By quantifying the percentage of live cells after exposure to a novel drug, researchers are able to determine therapeutic windows.[11] In regenerative medicine, tracking morphological changes and migration of stem cells is vital for optimizing tissue engineering strategies.[12,13] Segmenting individual stem cells within a scaffold helps scientists monitor their differentiation and spatial organization, which are key factors in creating functional tissues.[14,15] Segmentation also enables the quantification of metrics such as cell morphology, proliferation rates, and viability, which are crucial for understanding disease progression, evaluating treatment efficacy, and monitoring cellular health.[16–20] Live-dead segmentation tasks can be particularly challenging in three-dimensional tissue analysis,[21] and this had become one of the most active directions in the in-vitro studies recently to address heterogeneous cellular systems in cancer, metabolic, neurodegenerative, and infectious disease. These biomedical needs demonstrate that accurate segmentation is not merely a technical challenge but a foundational requirement for advancing biomedical research, enabling a better understanding of cellular behavior and driving innovations in healthcare solutions.

Quantitative phase imaging (QPI)[22] is a family of advanced techniques capable of imaging cell cultures or thick 3D cellular constructs noninvasively. It provides high-resolution, label-free visualization of cellular structures by quantifying the optical phase shift caused by nanometer path-length variations in the topography, dry mass, and refractive index of the sample. In

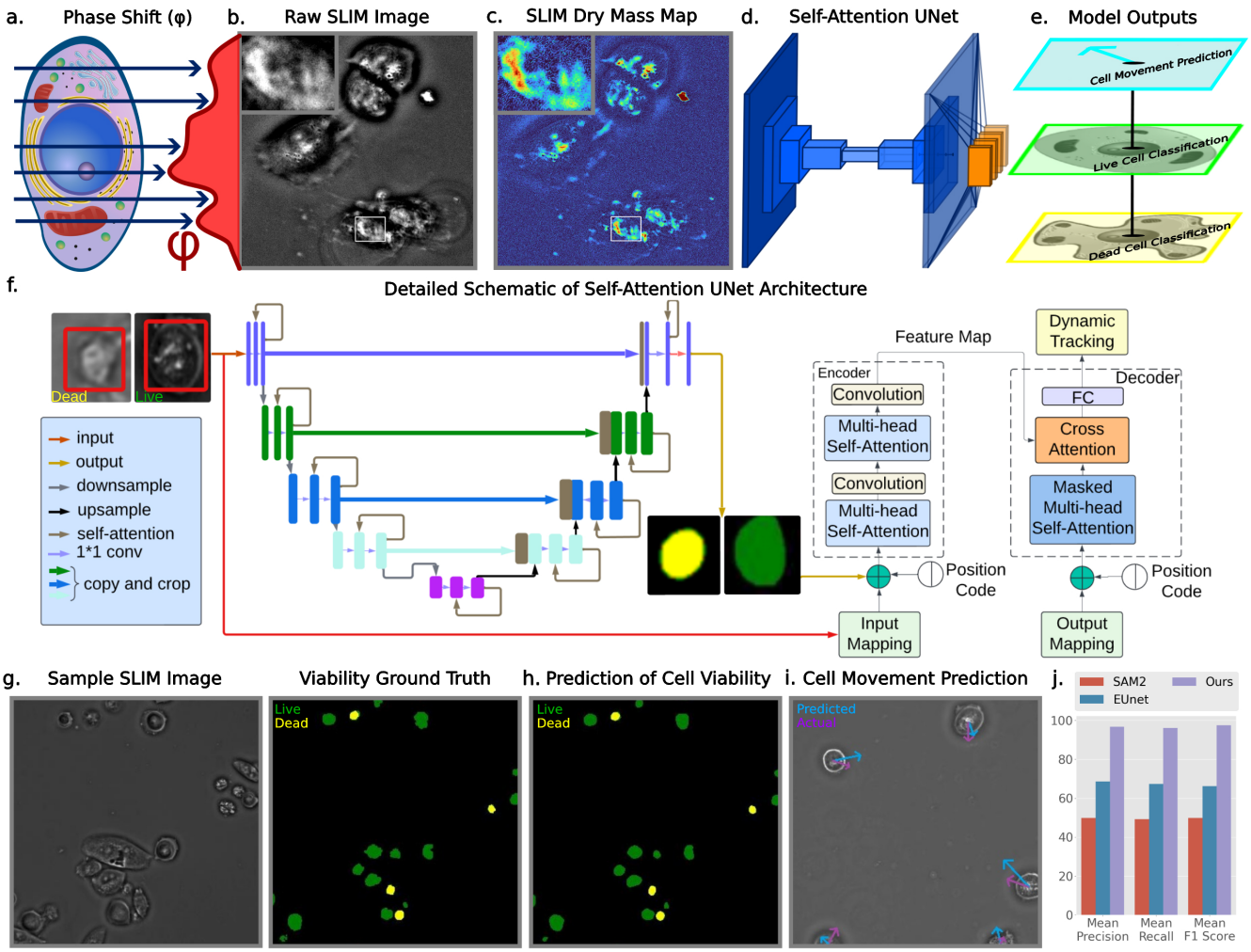

Figure 1. Overview of the proposed artificial intelligence (AI)-based architecture for cell viability segmentation and cell dynamics analysis. a. The phase of non-coherent light (dark blue arrows) is altered when passing through biological media. Phase shift (ϕ) is quantitatively measured by Spatial Light Interference Microscopy (SLIM - Phi Optics, Inc.). **b.** Sample QPI image of unlabeled CHO cells captured by SLIM. **c.** A dry mass heatmap generated from the same SLIM image exemplifies quantitative biological information imbued in QPI images by SLIM. **d.** Schematic of self-attention UNet architecture used to process raw SLIM images of unlabeled cells. Notably, a transformer block consisting of multi-head self-attention blocks is appended to the UNet output. **e.**Diagram of outputs made possible by the model used in this study, including cellular movement and cell viability. **f.** Detailed schematic of self-attention UNet architecture. In red squares, dead (left) and live (right) training data are fed into a UNet containing self-attention layers in every convolutional block. Input data and output segmentation maps are fed into the transformer appended to the end of the UNet, enabling cell movement prediction. **g.** Sample 2048 × 2048 SLIM image and associated manual annotation. **h.** Predicted viability of sample image by self-attention UNet. **i.** Visualization of virtual cell movement ground truth and AI predictions in a sample image. **j.** Comparison of mean viability segmentation performance across studied datasets using common and previously used machine learning models, with the model proposed in this study achieving demonstrably higher performance.

this paper, we use spatial light interference microscopy (SLIM), a highly sensitive implementation of QPI, which combines interferometry with white light holography that removes speckles present in laser-based QPI techniques and improves optical sectioning.[23] SLIM is implemented as an add-on to an optical microscope, so QPI and fluorescence channels are captured with pixel-to-pixel registration, which greatly simplifies machine learning training tasks. Using pairs of co-localized SLIM and fluorescence images, an ML algorithm can be trained to predict with high accuracy the location and extent of the same fluorescence marker in non-tagged samples (i.e., digital staining) in a process dubbed Phase Imaging with Computational Specificity (PICS).[24] PICS using UNet models has been demonstrated in various cell assays, including viability,[25] cell cycle phases,[26] viral detection,[27] lipid droplet classification,[28] and more. Effective segmentation and classification techniques play a

pivotal role in extracting meaningful information.

Segmentation can be approached using either non-deep learning or deep learning techniques. Current approaches face significant biological and clinical challenges. Traditional non-deep learning methods, such as thresholding and edge detection, rely on manual annotation or classical image processing algorithms.[29,30] While these methods are computationally efficient, they often fail to capture subtle differences in optical properties, such as the variations between live and dead cells in phase contrast images.[31,32] This drawback is particularly pronounced in high-throughput environments, where the cost of failure can be substantial. Misclassification of live-dead cells can lead to incorrect conclusions in drug efficacy studies or compromised treatment planning in oncology. These limitations, compounded by their labor-intensive nature, make them less effective for complex tasks like image segmentation.[23] On the other hand, deep learning approaches, particularly convolutional neural networks (CNNs),[33] have revolutionized segmentation by automating feature extraction and providing superior accuracy. Computational algorithms and machine learning models, such as UNet,[34] excel in biomedical image analysis by balancing precise localization with the preservation of spatial context. The UNet architecture's ability to efficiently capture spatial features across different receptive fields makes it ideal for biomedical image segmentation.[35] A major limitation of UNet is its inability to capture global features[36,37], which hampers its ability to segment entire images with high precision. Additionally, transformer-based architectures[38], including Vision Transformers (ViTs)[39] and Swin Transformers[40], leverage self-attention mechanisms to capture long-range dependencies across entire images, offering an edge in densely packed or morphologically diverse datasets[41]. However, the high computational cost of transformer models remains a challenge[42], underscoring the trade-off between performance and resource efficiency in selecting segmentation techniques. Moreover, transformer architectures lack inductive bias, particularly locality,[43,44], which can make them less effective at modeling fine-grained, spatially coherent features.

To address these limitations, we integrate a self-attention mechanism into the UNet architecture, effectively combining the strengths of UNet's localized feature extraction with the global feature modeling capabilities of transformers. This hybrid approach not only enhances segmentation accuracy but also enables the classification of cell viability (live versus dead), a critical aspect of biological research and clinical applications. Furthermore, our model leverages transformer architecture to predict future cell movements, providing a robust and comprehensive solution for the challenges posed by assay cell segmentation and dynamic cellular analysis.

Specifically, our contributions are as follows:

- We introduce a novel self-attention UNet model that builds upon the strengths of the standard UNet architecture by incorporating attention mechanisms to enhance the model's ability to differentiate between live and dead cells in SLIM images. The self-attention mechanism allows the model to focus on important features across the image, addressing the challenge of subtle phase-contrast variations that are difficult for traditional methods to detect.

- To extend the utility of our segmentation model, we integrate it with a transformer network, which extracts the global information from the entire input features. This allows the model to predict the future positions of cells based on their segmented states in sequential time-lapse images. By leveraging the long-range temporal dependencies captured by the transformer network, our model can anticipate cell movements, providing valuable insights into dynamic processes such as cell migration, proliferation, and morphological changes.

- Our combined approach offers a dual benefit: high-precision segmentation of cells (with a focus on live-dead classification) and predictive modeling of cell movement over time. This dual capability makes our framework especially valuable for long-term biological studies, where understanding both the static and dynamic aspects of cell populations is crucial. The integrated system surpasses traditional methods by not only segmenting static images but also by enabling predictive analysis, which can aid in tracking cellular responses to treatments or environmental changes.

2 Results

We developed a deep learning (DL) architecture specifically designed for SLIM images to enhance cell segmentation and cell motion forecasting (Fig. 1f). Our DL model incorporates self-attention mechanisms within a UNet framework to distinguish live from dead cells in complex imaging environments. The UNet architecture combines convolutional layers, self-attention mechanisms, and skip connections to extract spatial features. The use of self-attention enables the network to capture long-range dependencies within images, significantly improving segmentation and viability classification performance. To further extend this DL architecture, we integrated it with a transformer-based dynamic tracking module. This module employs an encoder-decoder transformer with multi-head self-attention and cross-attention mechanisms, incorporating position codes in the input and output mappings to preserve spatial and temporal coherence. By leveraging this combined design, we enabled accurate prediction of future cell positions using sequential imaging data.

In addition to segmentation and tracking, we analyzed cell movement patterns by computing spatial and temporal velocity correlations and displacement moments (Fig. 2a-d). Cell velocities were derived from sequential SLIM images by segmenting live and dead cells into binary masks and tracking their positions across time frames. Spatial correlations were computed by averaging velocity relationships across bins of spatial separation (ΔR), while temporal correlations were calculated by averaging velocity persistence across temporal lags (Δt). Displacement moments were evaluated for $q = 1, 1.5, 2, 2.5, 3$, and the scaling function $\lambda(q)$ was extracted using log-log plots. Quantifying these velocity correlations and displacement moments is important for understanding how cells coordinate movements under different conditions, respond to local stimuli, and shape tissue-level organization. Such data shed light on collective cell behaviors relevant to processes like wound repair and tumor invasion, where coordinated motion and rearrangements can influence overall outcomes. Both quantitative and qualitative evaluations demonstrated that our approach achieves superior segmentation accuracy, reliable movement prediction, and detailed characterization of cell dynamics, effectively addressing critical challenges in SLIM-based cellular analysis.

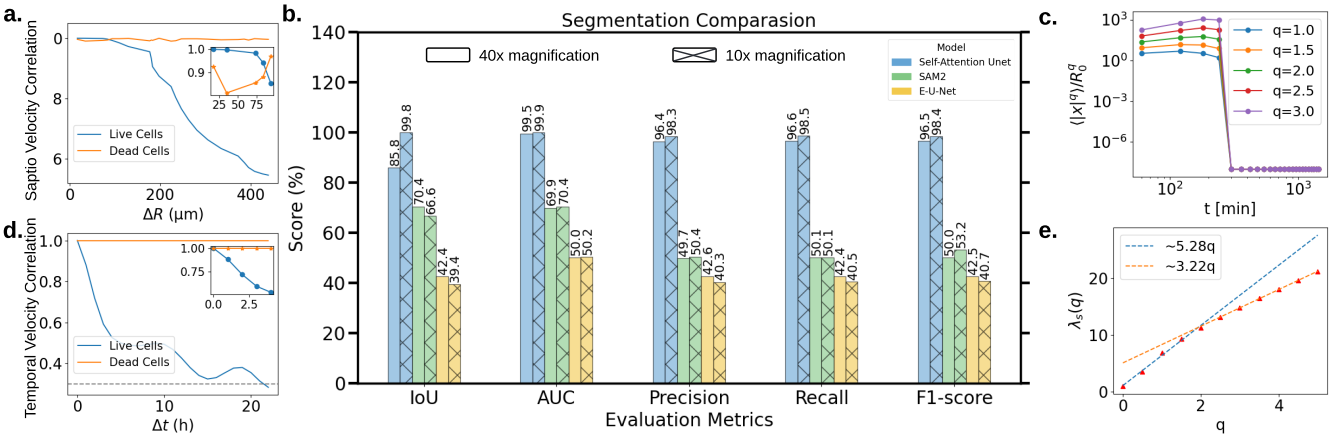

Figure 2. Performance analysis of the proposed Self-Attention UNet and other DL models. a. Spatial velocity correlation of live and dead cells in our dataset, showing the average radial distances (AR) for both cell types. **b.** Segmentation performance comparison across different models, including Self-Attention UNet, SAM2, and E-U-Net, using IoU, AUC, precision, recall, and F1-score metrics on 40× and 10× magnification images. All models utilize the same optimization function and loss function. **c.** Log-log plot of the average absolute displacement moments $\langle |x|^q \rangle / R_0^q$, with $R_0 = 2.07 \mu m$ representing the average maximum displacement of CHO cells. **d.** Temporal velocity correlation of live and dead cells, depicting the temporal variation of cell movement over time. **e.** $\lambda_S(q)$ as a function of q. The red triangles represent empirical values derived from tracking data over time intervals. The dashed blue and orange lines indicate the bi-linear fit slopes for low and high moments, respectively, with the slopes shown in the legend.

2.1 Cell Segmentation and Viability Classification

Our segmentation model processes SLIM images to detect and delineate cells while simultaneously classifying their viability using combined live and dead cell masks. To optimize performance, we employed a combination of cross-entropy and Jacard losses. The cross-entropy loss ensures precise pixel-wise classification, which is essential for distinguishing live and dead cells, while Jacard loss improves the overall segmentation quality by aligning predicted and ground-truth masks. The model was evaluated across 40× and 10× magnifications, addressing challenges related to varying cell size, shape, and fusion status. Representative results are presented in Fig. 2e; our DL model consistently demonstrates robust performance, effectively segmenting individual cells and reliably distinguishing between live and dead states, achieving high performance metrics including Intersection over Union (IoU), Area Under Curve (AUC), Precision, Recall, and F1-score across magnifications.
For cells at 40× magnification, which included larger, fused cells, the model achieved an IoU score of 85.8% and an AUC score of 99.5%. For cells at 10× magnification, which showed more typical morphologies, IoU scores exceeded 99.5%, and AUC scores reached 99.9%, demonstrating the model's robustness and ability to generalize across varied cell populations. To evaluate our AI model at an object-based level, we use metrics such as F1, Recall, and Precision. The results of the confusion matrix shown in Tab. 1 demonstrate the high accuracy of the trained model in classifying live and dead CHO cells. Specifically, the model correctly identified **99.9%** of the live cells, while only 0.1% of live cells were misclassified as dead. Similarly, **99.9%** of the dead cells were correctly identified, with only 0.1% of dead cells misclassified as live. The precision, recall, and F1 scores for both live and dead cells are all above 98%, indicating a balanced performance in terms of accuracy, sensitivity, and overall predictive power. These results suggest that the model is highly effective in differentiating between live and dead cells, with minimal misclassification errors, and maintains high confidence in its predictions. The high precision and recall

values further indicate that the model has low false-positive and high true-positive rates for both classes.

		Ground truth	
		Live ($n = 285002$)	Dead ($n = 47331$)
Cells	Live	99.9%	0.1%
	Dead	0.1%	99.9%
Evaluation	Precision	98.3%	98.6%
	Recall	98.5%	98.8%
	F1 Score	98.4%	98.8%

Table 1. Pixel-wise prediction and evaluation metrics for live and dead CHO cells. The trained model achieves high precision, recall, and F1 scores in distinguishing live and dead cells, with performance closely aligned with the ground truth. Confusion matrix entries are normalized relative to the number of cells in each class, reflecting the model's ability to accurately segment nuclear regions and classify cell viability.

To benchmark the proposed model, we compared it against two baseline models: the Segment Anything Model (SAM2)[45] and E-U-Net[25]. SAM2 is a general-purpose segmentation framework designed for diverse applications, offering adaptability across a wide range of datasets by leveraging prompt-based annotations and a robust backbone architecture. In contrast, E-U-Net is specifically tailored for biomedical imaging tasks and incorporates an encoder-decoder structure with skip connections to capture fine-grained details in cell images. All models were evaluated on the same dataset under identical conditions to ensure a fair comparison.

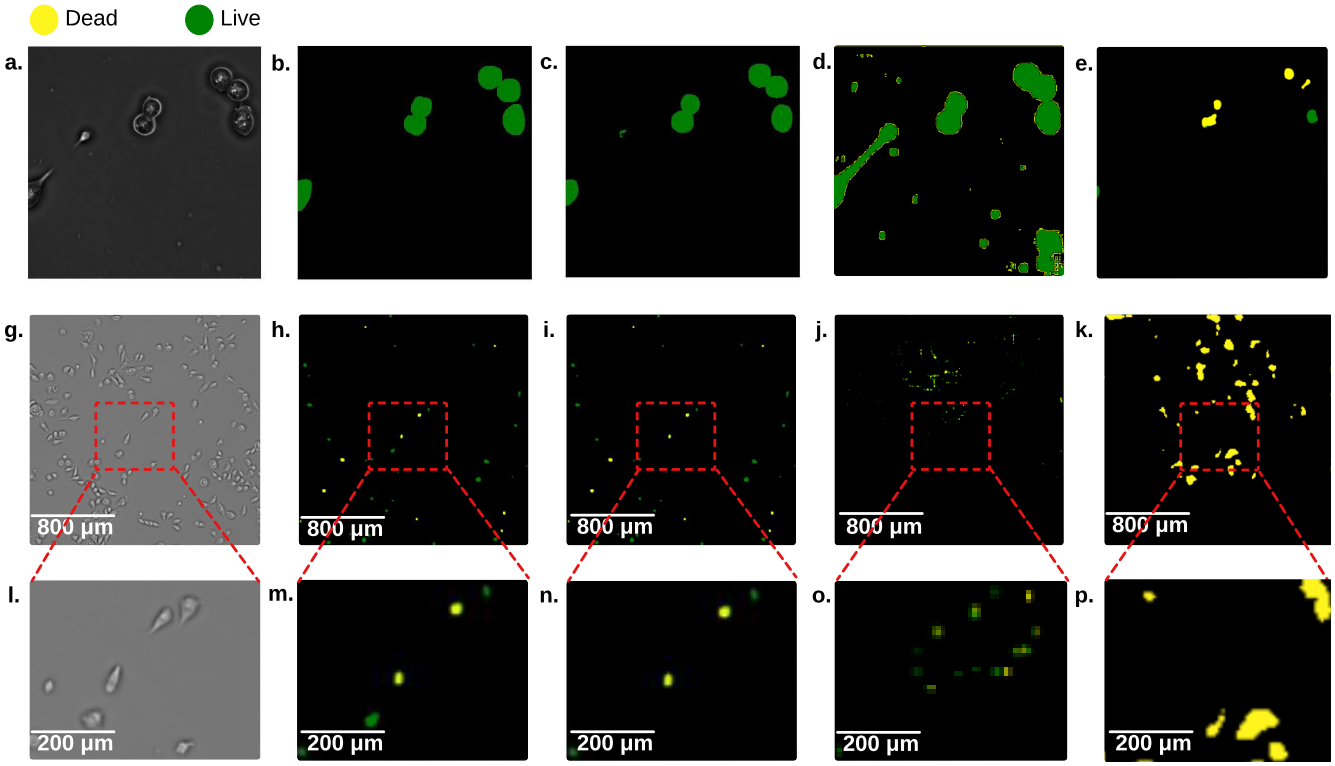

Figure 3. Model performance on the validation dataset. Representative results of the model on validation SLIM images under 40× and 10× magnifications. **a.** Input SLIM image at 40× magnification. **b.** Ground truth for a., showing cell viability states: background (black), live (green), and dead (yellow). **c.** Segmentation result using the proposed AI model for a. **d.** Segmentation result using SAM2 for a. **e.** Segmentation result using E-U-Net for a. **g.** Input SLIM image at 10× magnification. **h.** Ground truth for g. **i.** Segmentation result using the proposed model for g. **j.** Segmentation result using SAM2 for g. **k.** Segmentation result using E-U-Net for g. **l.** Zoomed-in region of interest from g. at 200 μm scale. **m.** Corresponding zoomed-in region for h. **n.** Corresponding zoomed-in region for i. **o.** Corresponding zoomed-in region for j. **p.** Corresponding zoomed-in region for k.

Interestingly, both SAM2 and E-U-Net struggle with segmenting cells with 10× magnification, which often lack detailed features(Fig. 3). In contrast, the self-attention Unet model could effectively segment small and large cells, achieving high IoU scores. Notably, our model outperforms SAM by over 6% in segmenting 40× magnification cells and by nearly 50% in 10× magnification cells, along with outperforming E-U-Net across different magnifications. This improvement is attributed to the self-attention mechanism, which enhances feature extraction in low-resolution datasets where SAM and E-U-Net struggle with detailed segmentation. We also performed a direct comparison on all three models of both our dataset and the E-U-Net dataset to examine the generality of our model. The results in Table 2 show that our model, which was trained on our dataset, achieved considerably better performance metrics than E-U-Net on the same dataset. To ensure fairness, we also tested our model on the E-U-Net dataset, where our model surpassed E-U-Net by nearly **5%** in terms of the F1 score while maintaining similar Recall and Precision. On both datasets, our model has surpassed SAM2 in a wide range (larger than 45%).

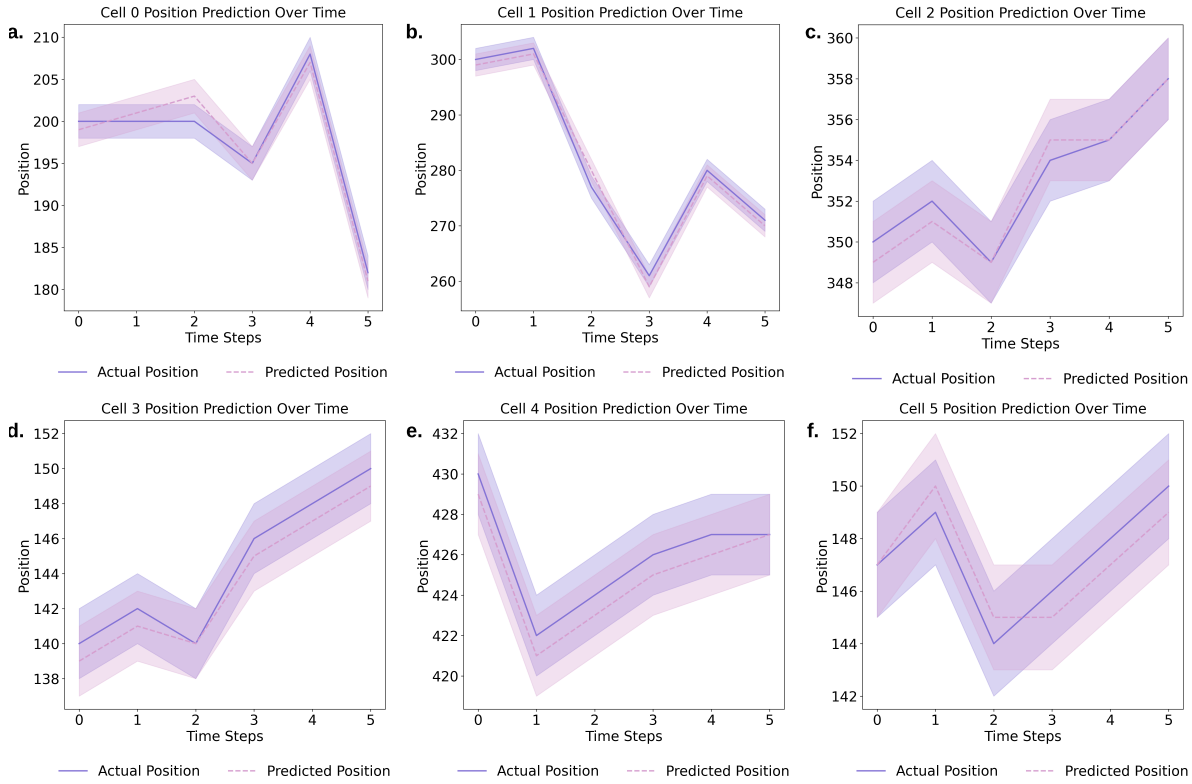

Figure 4. Predicted and actual cell positions over time for a randomly selected image. The plots represent the time evolution of actual positions (solid lines) and predicted positions (dashed lines) for six cells identified in a randomly selected image. The x-axis denotes time steps, and the y-axis indicates the position of each cell. Shaded regions around each curve correspond to a ±2.0 position variance, representing the uncertainty associated with actual and predicted positions. The subplots illustrate individual cell trajectories, showing overall alignment between predictions and actual positions, with minor deviations observed in certain cases.

2.2 Dynamic Cell Tracking and Spatio-temporal movement forecasting

The transformer network was trained to predict cell positions over five time steps, representing 1-hour intervals, using sequences of prior cell states derived from the UNet model's segmentation outputs. This sequence length balances computational efficiency and the need to capture short-term temporal dependencies critical for dynamic cell analysis.Mean squared error (MSE) loss was used during training to minimize the discrepancy between predicted and actual positions. Here, the position is defined as the Euclidean distance from the origin, calculated as $\sqrt{x^2+y^2}$, where x and y represent the coordinates of the cell. MSE loss is particularly suitable for this task as it penalizes larger deviations more heavily, encouraging the model to make precise predictions of cell trajectories. The performance of the network was evaluated using Euclidean distance, which measures the spatial error between predicted and ground truth positions. Euclidean distance provides an interpretable metric for quantifying prediction accuracy, with smaller values indicating closer alignment between predicted and actual cell trajectories. Together, these metrics demonstrated the ability of the transformer network to predict cell movements over time reliably.

As shown in Fig. 4, the predicted and actual positions of cells aligned closely across most samples, validating the model's effectiveness in tracking cellular movements over time. To further evaluate the model's predictions, we compared the predicted and actual positions for every cell across all samples. The heatmaps in Fig. 5 illustrate the distribution of Euclidean distances between predicted and actual positions, measured in the unit of $\mu m/pixel$, with the majority of samples showing small errors within a range of less than 6 units. Aggregated predictions for one time step across the entire validation set further corroborate the accuracy of the model, with an overall loss of less than 8.39×10^{-5}, underscoring its precision in predicting cell movements. Figure 6e demonstrates how predicted cell trajectories closely track actual observed movements, with only minor deviations in some instances. Analyzing the distribution of prediction errors reveals that most predictions fall within a small error margin of 0 to 5 units. These findings confirm the model's capability to reliably predict cell movements over time.

Dataset	Model	Metrics (%)		
		Precision	F1 Score	Recall
Our Dataset	E-U-Net	42.6	42.5	42.4
	SAM2	49.7	50.0	50.1
	Our Model	98.3	98.5	98.4
E-U-Net Dataset	E-U-Net	94.6	90.1	92.3
	SAM2	50.1	49.8	48.4
	Our Model	95.1	96.4	93.7

Table 2. Comparison of metrics between E-U-Net, SAM2, and the proposed AI model on two datasets. The proposed AI model significantly outperforms E-U-Net and SAM2 across precision, F1 score, and recall metrics on both the model's dataset and the E-U-Net dataset. These results highlight the robustness and adaptability of the proposed model in accurately segmenting and classifying cell states across different datasets and magnifications.

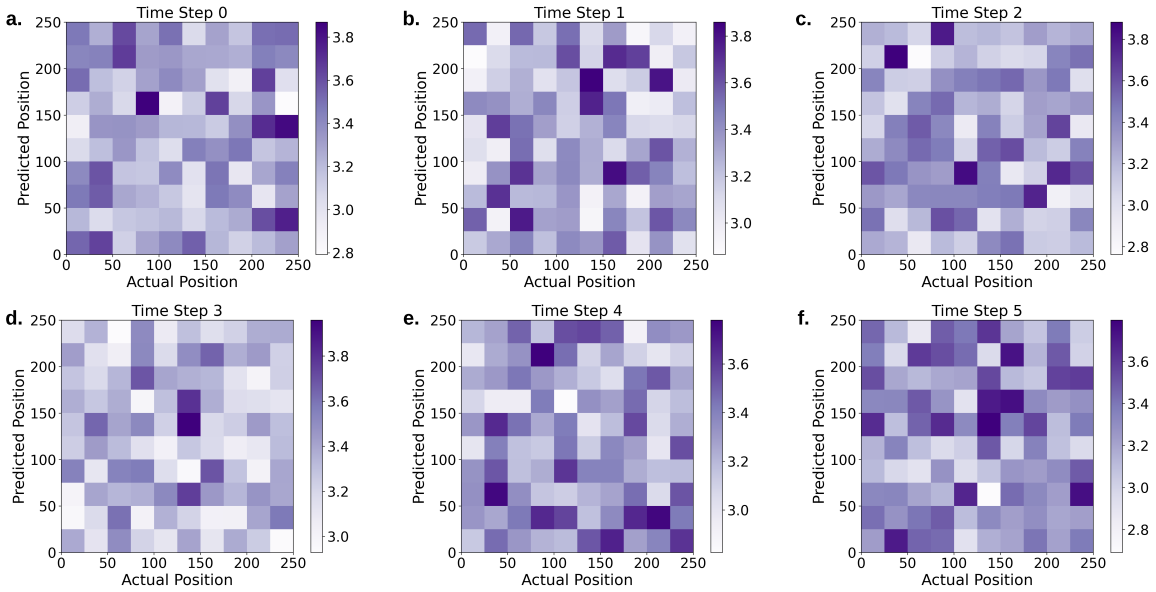

Figure 5. Heatmaps of Euclidean distance between actual and predicted positions for all training images across time steps 0 to 5(1 hour per time step). Each heatmap represents the Euclidean distance between actual positions (x-axis) and predicted positions (y-axis) for all training images at a specific time step. The intensity of the heatmap color indicates the magnitude of the Euclidean distance, with darker shades representing larger distances. These visualizations provide insight into the distribution and variability of prediction errors across different time steps.

In addition to predicting individual cell trajectories, the model successfully forecasted the overall distribution of live and dead cell states across entire images for two future time steps. The current cell state mask is shown in Fig. 6a, with a zoomed-in region highlighted in Fig. 6b. The ground truth masks for the first and second time-steps are depicted in Fig. 6c and Fig. 6d, respectively, while the corresponding predictions from the model are presented in Fig. 6f and Fig. 6g. During these two time-steps, dead cells remained stationary while live cells moved. For the first time step, the model accurately predicted the trajectories of live cells, closely aligning with the ground truth. However, for the second time step, some discrepancies were observed, with the predicted positions of a few live cells deviating from the ground truth. Despite these minor misalignments, the model achieved an IoU of 0.65, demonstrating reliable accuracy in capturing cell viability and movement over short time periods. These results, illustrated through sample images, underscore the model's potential for studying cellular dynamics, monitoring responses to external stimuli, and analyzing disease progression in biological systems.

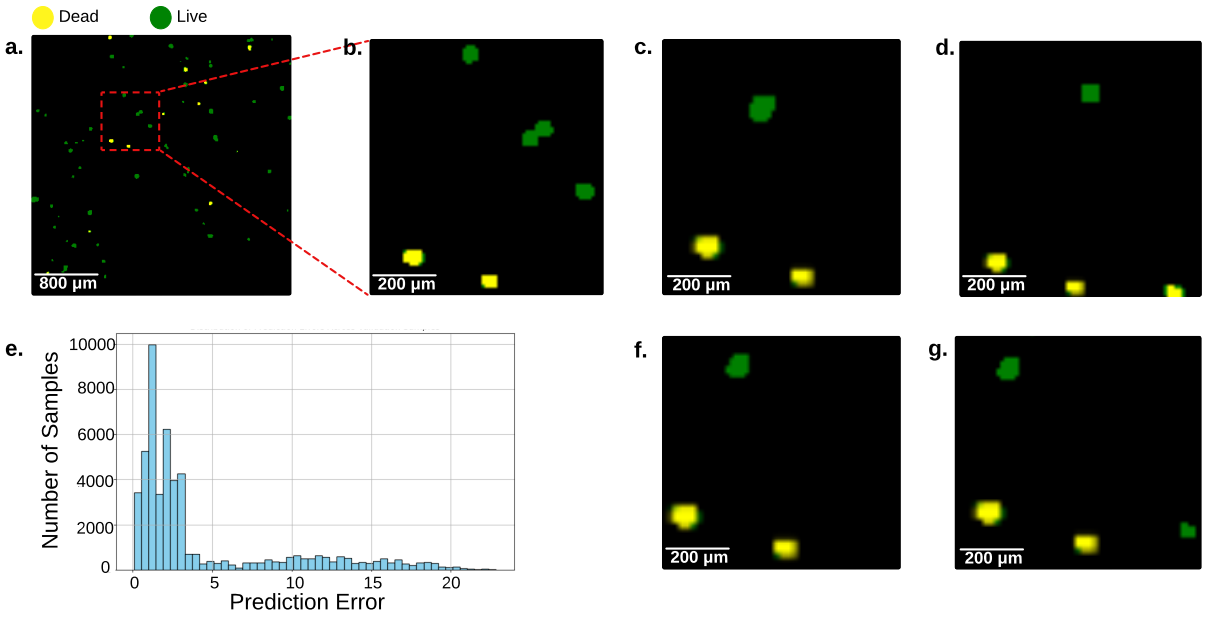

Figure 6. Visualization of predicted and ground truth cell states and prediction errors. a. Current image mask showing live cells (green) and dead cells (yellow). **b.** Zoomed-in region of interest from **a** at a 200 µm scale. **c.** Ground truth mask of cell movement for the first future time step. **d.** Ground truth mask for the second future time step. **e.** Histogram illustrating the distribution of Euclidean distance errors between actual and predicted positions for all validation samples of one time step, with the majority of errors falling below 5 units. **f.** Predicted mask by the proposed model for the first future time step. **g.** Predicted mask by the proposed model for the second future time step.

3 Discussion

Traditional integration of neural networks into image analysis workflows focuses on solving a single task, such as segmentation, classification, or forecasting. In this study, we suggest that AI research need not limit their models to a single function. We developed an advanced AI framework combining UNet-based semantic segmentation with a transformer network to both predict cell viability and forecast cell movements, thereby providing a powerful tool for analyzing cellular dynamics in live imaging data. Beyond the novelty of a combined approach, our work demonstrates significant improvements in both segmentation accuracy and movement prediction, addressing key challenges in distinguishing live and dead cells in complex cellular environments.

The self-attention UNet model proved highly effective in distinguishing between live and dead cells across various cell magnifications. The introduction of self-attention mechanisms into the UNet architecture allowed the model to capture long-range dependencies, which potentiates learning for biological patterns present across spatial dimensions too large to be processed by traditional convolution and could explain our model's notable improvement in performance over the traditional UNet model. Specifically, the model achieved IoU scores exceeding 0.97 across magnification datasets. Not only does this surpass the performance of traditional segmentation approaches, but it also suggests our model's robustness when faced with diverse biological and morphological inputs. This is essential for any model seeking widespread adoption in biological research, where high-density or morphologically complex environments are increasingly common.

Beyond segmentation, our integration of the transformer network at the output of the UNet represents a novel approach to cell movement forecasting from sequential time-lapse imaging data. The transformer network, designed for handling long-term dependencies, was able to predict future cell positions with high precision, as evidenced by the minimal prediction error observed across validation samples. Most prediction deviations fell within 0 to 5 units of Euclidean distance error, demonstrating the model's suitability for tasks requiring precise tracking, such as monitoring cell migration or proliferation in response to treatment. By integrating segmentation, dynamic tracking, and movement prediction, our framework provides a comprehensive tool for analyzing and monitoring individual cell behaviors over time.

Despite these strengths, some limitations should also be acknowledged. The model's performance decreases slightly in cases with highly irregular cell morphologies or overlapping cells, where segmentation and tracking tasks become more challenging. Additionally, the computational demands of the transformer network may limit scalability for large datasets or real-time applications, particularly in high-throughput settings.

The self-attention UNet-Transformer integrated framework not only enhances segmentation and prediction capabilities but also offers significant potential for broader applications in biomedical research and clinical settings. Future work could focus on optimizing computational efficiency for real-time analysis in high-throughput settings. Incorporating multi-modal imaging (e.g., fluorescence and QPI) could enhance robustness and generalizability. Expanding the framework to classify additional cell states or integrate them into clinical workflows may further bridge the gap between research and application. Adapting the model for real-time analysis and evaluating its performance across diverse imaging modalities, including multi-dimensional datasets, could further broaden its utility and impact.

Future investigation could also be performed on model latent space to determine which cellular features revealed by quantitative phase imaging are used to predict viability. Furthermore, a latent transformer model could be used to determine which cellular features are used by the final transformer node to predict cellular movements. These advancements would bridge the gap between the black box, which is computer vision models, and the underlying biological processes models used to predict cellular states and movements. Our model's unique combination of cell viability classification and cellular movement prediction has diverse applications in wound healing assays, as well as immune response. The cell health and movement components of this method also potentiate diverse studies in oncology and pharmacology. Our model could enable predictive studies of tumor resistance and cancer relapse as well as real-time monitoring of cellular responses to treatment. Ultimately, the myriad studies enabled by the combination of cellular viability assessment and movement prediction point to the power of machine-learning approaches with multiple readouts and suggest that these methods could significantly advance bridging the gap between computational modeling and clinical translation.

4 Methods

4.1 Data Preparation

Images were acquired with a SLIM module (Phi Optics, USA) attached to the camera port of a motorized phase-contrast microscope (Carl Zeiss, Germany, and Leica, Germany) equipped with four additional fluorescence channels complementing SLIM structural information, with the ability to perform correlative imaging to study cellular constituents with molecular specificity

The experimental platform integrates a calibrated mini-x-ray source with the QPI/microscopy system with access to short to long-timescale observation windows to potentiate effective radiosensitization interventions. The x-ray source mounted on the microscope is a custom Moxtek MAGNUM 50kVp XRF source employing a conduction-cooled transmission W target operating at 4-50 kVp and 0-200 µA operating at a maximum power of 4W. The transmission target is 100 µm thick with a 125 µm Be window. Accelerated electrons strike the target with an effective spot size of 500 µm FWHM. The custom brass collimator with an aperture of 2 mm has a narrow snout of 2.7 cm length to accommodate the small space between the cell well and the microscope condenser. The experiments were done at 50 kVp/80 µA, with a dose output at 70cGy/min at the bottom of the cell well.

Chinese Hamster Ovary subclone K1 (CHO-K1) cells (ATCC CCL-61) were obtained from the American Type Culture Collection (ATCC) and maintained in the exponential growth phase. The cells were cultured in 60 mm tissue culture-treated dishes (Fisher Scientific) using F-12K medium (ATCC) supplemented with 10% fetal bovine serum (FBS) and penicillin/streptomycin (Gibco).

For experiments, 100,000 CHO-K1 cells were seeded in 1.5 mL of F-12K complete medium in 35 mm tissue culture-treated glass-bottom dishes. Cells were allowed to adhere and grow for 16 hours prior to treatment. One drop of each reagent from the ReadyProbes™ Cell Viability Imaging Blue/Green Kit (Invitrogen, USA) was added to the culture medium 20 minutes before radiation and live image acquisition. The kit consists of DNA nuclear dyes as fluorescence markers for ground truth training: NucBlue (the "live" reagent) combines with the nuclei of all cells and can be imaged with a DAPI fluorescent filter set, and NucGreen (the "dead" reagent) stains the nuclei of cells with compromised membrane integrity, which is imaged with a FITC filter set.

Imaging was done using 10×/0.30 and 40×/0.75 phase objectives with the cells incubated at 37C/5% CO2. Baseline imaging was done prior to irradiation using sequential imaging as follows: SLIM channel / GFP channel/ DAPI channel for each tile of a 10×10 mosaic ($4mm^2$ total area). The cells were irradiated for 5 minutes, to a total dose of 350cGy, while imaging was performed on one field-of-view tile, live on all 3 channels, at 20s intervals, for a total time of 20 minutes. Subsequent scans were set to run overnight using the same scanning formula as the baseline (10×10 mosaic), with 18 repetitions at 1-hour intervals. A second irradiation of 490cGy was performed the next day, 20 hours after the 1st irradiation, under the same scanning conditions.

4.2 Self-Attention UNet-Transformer Architecture

The Self-Attention UNet model is a UNet-based model with self-attention added to each layer. We integrated a multi-head self-attention mechanism[46]into the UNet architecture to capture long-range spatial dependencies and enhance feature learning for image segmentation.

We first reshape our input feature map X to attention-desired size $X' \in \mathbb{R}^{B\times H\times W\times C}$, where B is the batch size, $H \times W$ represents the spatial dimensions, C is the number of channels. Then, applied multi-head self-attention with four attention heads. Attention weights are computed using the scaled dot-product attention mechanism:

$$Attention(Q,K,V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

where $Q = X'W_Q, K = X'W_K, V = X'W_V$, and d_k is the dimensionality of the queries and keys. This operation allows the network to weigh the importance of different spatial locations, effectively modeling long-range dependencies.

After that, the output of the attention module is combined with the input via a residual connection, followed by a feed-forward network consisting of two fully connected layers with GELU (Gaussian Error Linear Unit) activation[47]:

$$A_{out} = GELU(AW_1 + b_1)W_2 + b_2 + A \tag{2}$$

where A is the output of the attention mechanism, and W_1, W_2, b_1, b_2 are all learned parameters.This enables the network to preserve the original features while incorporating attention-based enhancements.

To model and predict the movement of cells over time, we integrated a transformer network into the pipeline. The input to the transformer model consists of sequences of tracked cell positions over time extracted from the UNet segmentation masks. We trained the transformer to predict future cell positions based on past sequences. We tracked the positions of live and dead cells across time frames by extracting the cell coordinates from the segmented masks. For each sequence of time steps, we generated cell tracks and used these tracks as input to the transformer model. The image dimensions normalized the positions to ensure consistent training.

The transformer architecture includes an embedding layer to project the 2D cell positions into a higher-dimensional space, followed by a multi-layer Transformer encoder. The Transformer encoder consists of two layers, each with four attention heads, which allow the model to capture temporal dependencies in the cell positions.

The output of the Transformer encoder is passed through a linear decoder layer to predict the future positions of the cells.

4.3 Spatial and Temporal Velocity Correlations and Displacement Moments

To analyze the movement dynamics of live and dead cells, we computed spatial and temporal velocity correlations[48] alongside displacement moments[49]. These metrics provide insights into coordinated movement patterns, persistence over time, and scaling behaviors in cellular dynamics.

4.3.1 Spatial Velocity Correlation

The spatial velocity correlation quantifies how the velocities of cell pairs are related as a function of their spatial separation (ΔR):

$$C_v(\Delta R) = \frac{\sum_{i,j} \mathbf{v}_i \cdot \mathbf{v}_j \, \delta(|\mathbf{r}_i - \mathbf{r}_j| - \Delta R)}{\sum_{i,j} \delta(|\mathbf{r}_i - \mathbf{r}_j| - \Delta R)}, \tag{3}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ are the velocity vectors of cells i and j, $\mathbf{r}_i$ and $\mathbf{r}_j$ denote their respective positions, and δ is the Dirac delta function. This metric evaluates how movement coordination varies with distance.

4.3.2 Temporal Velocity Correlation

The temporal velocity correlation measures the persistence of cell movement over time intervals (Δt):

$$C_v(\Delta t) = \frac{\langle \mathbf{v}(t) \cdot \mathbf{v}(t+\Delta t) \rangle}{\langle |\mathbf{v}(t)|^2 \rangle}, \tag{4}$$

where $\mathbf{v}(t)$ is the velocity vector of a cell at time t. This metric captures the consistency of cell movement over varying temporal lags.

4.3.3 Displacement Moments

Displacement moments of order q characterize the scaling behavior of cell displacements and were computed as:

$$\langle |x(t)|^q \rangle = \frac{1}{N} \sum_{i=1}^{N} |x_i(t) - x_i(0)|^q, \tag{5}$$

where $x_i(t)$ and $x_i(0)$ are the positions of cell i at time t and the initial time, respectively, and N is the total number of cells. The moments were normalized using the average maximum displacement (R_0):

$$R_0 = \frac{1}{N} \sum_{k=1}^{N} \max_{i,j} \|\mathbf{r}_i - \mathbf{r}_j\|, \tag{6}$$

where R_0 is derived as the mean of the largest pairwise displacements ($R_{\max}$) within each cell track, this normalization allows direct comparison across different timescales and magnifications. Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ denote cells' respective positions

4.3.4 Moment Scaling Function

The scaling behavior of displacement moments was further analyzed using the moment scaling function $\lambda(q)$:

$$\lambda(q) = \frac{\log \langle |x(t)|^q \rangle}{\log t}. \tag{7}$$

Distinct scaling regimes were observed, characterized by different slopes in $\lambda(q)$ for low and high moments (q), separated by a critical timescale of approximately 5 minutes.

4.4 Training parameters

We trained both the UNet and transformer models using the Adam optimizer with an initial learning rate of 0.001. For the UNet, we employed cross-entropy loss[50] to account for the three segmentation classes. Early stopping was implemented to prevent overfitting, with a patience of five epochs based on the loss value.

$$\mathscr{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \tag{8}$$

In the cross-entropy loss, $\hat{y}_i \in \mathbb{R}^C$ is the predicted probability distribution for pixel i over C classes (background, live cell, dead cell). $y_i \in \{0,1,2\}$ is the true class label for pixel i. $y_{i,c}$ is 1 if the true class of pixel i is c, and 0 otherwise. $\hat{y}_{i,c}$ is the predicted probability that pixel i belongs to class c.This loss ensures that the model maximizes the predicted probability for the correct class at each pixel.

For the transformer model, we split the cell position sequences into training and validation sets using a sequence length of five time steps. The model was trained for 100 epochs, and performance was evaluated using both the validation loss and the prediction error in Euclidean distance between predicted and true cell positions.

The mean squared error (MSE) loss[51] is used for the transformer model to predict future cell positions based on previous time steps. This loss compares the predicted coordinates of the cells with the actual coordinates.

$$\mathscr{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{y}_i - y_i\|^2 \tag{9}$$

In Eq. 9, $\hat{y}_i \in \mathbb{R}^2$ is the predicted (x,y) coordinates for cell i at the future time step. $y_i \in \mathbb{R}^2$ is the actual (x,y) coordinates for cell i. $\hat{y}_i$ is the predicted cell position. y_i is the true cell position. This loss penalizes large differences between the predicted and actual positions, ensuring accurate future cell position predictions.

We implemented the model using PyTorch. Training, validation, and testing were carried out on an NVIDIA Tesla V100 GPU with 54 GB of VRAM. After each training session, we saved the model weights and used them for transfer learning to the next dataset.

References

1. Edlund, C. *et al.* Livecell—a large-scale dataset for label-free live cell segmentation. *Nat. methods* **18**, 1038–1045 (2021).
2. Xing, F. & Yang, L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews biomedical engineering* **9**, 234–263 (2016).
3. Zimmer, C., Labruyere, E., Meas-Yedid, V., Guillén, N. & Olivo-Marin, J.-C. Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: A tool for cell-based drug testing. *IEEE transactions on medical imaging* **21**, 1212–1221 (2002).
4. Wen, T. *et al.* Review of research on the instance segmentation of cell images. *Comput. methods programs biomedicine* **227**, 107211 (2022).
5. Davatzikos, C. *et al.* Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *J. medical imaging* **5**, 011018–011018 (2018).
6. Bakas, S. *et al.* Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629* (2018).
7. Goyal, S. & Kataria, T. Image guidance in radiation therapy: techniques and applications. *Radiol. research practice* **2014**, 705604 (2014).
8. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. reviews Drug discovery* **18**, 463–477 (2019).
9. Esch, M., King, T. & Shuler, M. The role of body-on-a-chip devices in drug and toxicity studies. *Annu. review biomedical engineering* **13**, 55–72 (2011).
10. Gao, R. *et al.* Tapered chiral nanoparticles as broad-spectrum thermally stable antivirals for sars-cov-2 variants. *Proc. Natl. Acad. Sci.* **121**, e2310469121 (2024).
11. Savitski, M. M. *et al.* Tracking cancer drugs in living cells by thermal profiling of the proteome. *Science* **346**, 1255784 (2014).
12. Gomes, M. E., Rodrigues, M. T., Domingues, R. M. & Reis, R. L. Tissue engineering and regenerative medicine: new trends and directions—a year in review. *Tissue Eng. Part B: Rev.* **23**, 211–224 (2017).
13. Berthiaume, F., Maguire, T. J. & Yarmush, M. L. Tissue engineering and regenerative medicine: history, progress, and challenges. *Annu. review chemical biomolecular engineering* **2**, 403–430 (2011).
14. Dos Santos, A. X. d. S. & Liberali, P. From single cells to tissue self-organization. *The FEBS journal* **286**, 1495 (2018).
15. Loeffler, M. & Roeder, I. Tissue stem cells: definition, plasticity, heterogeneity, self-organization and models–a conceptual approach. *Cells Tissues Organs* **171**, 8–26 (2002).
16. Cutler, K. J. *et al.* Omnipose: a high-precision morphology-independent solution for bacterial cell segmentation. *Nat. methods* **19**, 1438–1448 (2022).
17. Hester, S. D., Belmonte, J. M., Gens, J. S., Clendenon, S. G. & Glazier, J. A. A multi-cell, multi-scale model of vertebrate segmentation and somite formation. *PLoS computational biology* **7**, e1002155 (2011).
18. Eren, F. *et al.* Deepcan: A modular deep learning system for automated cell counting and viability analysis. *IEEE journal biomedical health informatics* **26**, 5575–5583 (2022).
19. Alieva, M., Wezenaar, A. K., Wehrens, E. J. & Rios, A. C. Bridging live-cell imaging and next-generation cancer treatment. *Nat. Rev. Cancer* **23**, 731–745 (2023).
20. Guo, J. L., Januszyk, M. & Longaker, M. T. Machine learning in tissue engineering. *Tissue Eng. Part A* **29**, 2–19 (2023).
21. Lee, J., Lilly, G. D., Doty, R. C., Podsiadlo, P. & Kotov, N. A. In vitro toxicity testing of nanoparticles in 3d cell culture. *Small* **5**, 1213–1221 (2009).
22. Popescu, G. Quantitative phase imaging of cells and tissues. *J. Biomed. Opt.* **17**, 029901 (2012).
23. Chen, X., Kandel, M. E. & Popescu, G. Spatial light interference microscopy: principle and applications to biomedicine. *Adv. optics photonics* **13**, 353–425 (2021).
24. Kandel, M. E. *et al.* Phase imaging with computational specificity (pics) for measuring dry mass changes in sub-cellular compartments. *Nat. communications* **11**, 6256 (2020).
25. Hu, C. *et al.* Live-dead assay on unlabeled cells using phase imaging with computational specificity. *Nat. communications* **13**, 713 (2022).

26. He, Y. R. *et al.* Cell cycle stage classification using phase imaging with computational specificity. *ACS photonics* **9**, 1264–1273 (2022).

27. Goswami, N. *et al.* Label-free sars-cov-2 detection and classification using phase imaging with computational specificity. *Light. Sci. & Appl.* **10**, 176 (2021).

28. Sheneman, L., Stephanopoulos, G. & Vasdekis, A. E. Deep learning classification of lipid droplets in quantitative phase images. *PLoS One* **16**, e0249196 (2021).

29. Senthilkumaran, N. & Vaithegi, S. Image segmentation by using thresholding techniques for medical images. *Comput. Sci. & Eng. An Int. J.* **6**, 1–13 (2016).

30. Ghandorh, H. *et al.* Semantic segmentation and edge detection—approach to road detection in very high resolution satellite images. *Remote. Sens.* **14**, 613 (2022).

31. Zaburko, J., Staniszewski, M., Dziadosz, M., Babko, R. & Łagód, G. Automatic system for acquisition and analysis of microscopic digital images containing activated sludge. *Adv. Sci. Technol. Res. J.* (2024).

32. Zhang, D., Islam, M. M. & Lu, G. A review on automatic image annotation techniques. *Pattern Recognit.* **45**, 346–362 (2012).

33. Kayalibay, B., Jensen, G. & van der Smagt, P. Cnn-based segmentation of medical imaging data. *arXiv preprint arXiv:1701.03056* (2017).

34. Zeng, Z., Xie, W., Zhang, Y. & Lu, Y. Ric-unet: An improved neural network based on unet for nuclei segmentation in histology images. *Ieee Access* **7**, 21420–21428 (2019).

35. Guan, S., Khan, A. A., Sikdar, S. & Chitnis, P. V. Fully dense unet for 2-d sparse photoacoustic tomography artifact removal. *IEEE journal biomedical health informatics* **24**, 568–576 (2019).

36. Siddique, N., Paheding, S., Elkin, C. P. & Devabhaktuni, V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE access* **9**, 82031–82057 (2021).

37. Gu, Z. *et al.* Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging* **38**, 2281–2292 (2019).

38. Chen, J. *et al.* Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).

39. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

40. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022 (2021).

41. Wei, C., Ren, S., Guo, K., Hu, H. & Liang, J. High-resolution swin transformer for automatic medical image segmentation. *Sensors* **23**, 3420 (2023).

42. Han, K. *et al.* Transformer in transformer. *Adv. neural information processing systems* **34**, 15908–15919 (2021).

43. Xu, Y., Zhang, Q., Zhang, J. & Tao, D. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Adv. neural information processing systems* **34**, 28522–28535 (2021).

44. Bachmann, G., Anagnostidis, S. & Hofmann, T. Scaling mlps: A tale of inductive bias. *Adv. Neural Inf. Process. Syst.* **36** (2024).

45. Kirillov, A. *et al.* Segment anything. *arXiv:2304.02643* (2023).

46. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* (2017).

47. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).

48. Rabani, A., Ariel, G. & Be'er, A. Collective motion of spherical bacteria. *PloS one* **8**, e83760 (2013).

49. Vilk, O., Charter, M., Toledo, S., Barkai, E. & Nathan, R. Strong anomalous diffusion for free-ranging birds. *arXiv preprint arXiv:2411.04684* (2024).

50. Zhang, Z. & Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Adv. neural information processing systems* **31** (2018).

51. Hodson, T. O., Over, T. M. & Foks, S. S. Mean squared error, deconstructed. *J. Adv. Model. Earth Syst.* **13**, e2021MS002681 (2021).

5 Acknowledgements

Authors acknowledge useful discussions with Dr. Catalin Chiritescu. The research in this paper was supported by GRANT XXX, YYY, and ZZZ

6 Author contributions statement

C.B., D.I., N.K., S.N., and P.B. conceived the experiments, A.C., C.Y., A.S., and Y.C. conducted the experiments, M.L., M.S., M.D., and J.A. curated the data. All authors reviewed the manuscript.

7 Additional information

7.1 Code and Data Availability

Our code to run the experiments can be found at https://github.com/Belis0811/Unet.

7.2 More Experiments and Dicussion

7.2.1 Model Comparison and Metrics

Loss Functions. All models(self-attention Unet, SAM2, and E-U-Net) were trained using a combination of cross-entropy loss and Jaccard loss. Cross-entropy loss measures the difference between the predicted probability distribution and the true distribution. Jaccard loss penalizes mismatches between predicted and true masks, often used in segmentation tasks.

Evaluation Metrics. All models were evaluated on the same test set with identical criteria, allowing for a fair comparison. The primary metrics include:

- **Intersection over Union (IoU)**:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}},$$

  where TP, FP, FN stand for true positive, false positive, and false negative, respectively. IoU measures how much the predicted region overlaps with the ground truth.

- **Area Under the Curve (AUC)**: This is the area under the receiver operating characteristic (ROC) curve. It captures how well the model distinguishes between positive and negative examples over a range of classification thresholds.

- **F1-score**: The harmonic mean of precision and recall:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

  It emphasizes scenarios where both precision and recall are important.

- **Recall**:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

  This indicates the proportion of true positives that the model successfully identifies.

- **Precision**:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

  This indicates the proportion of predicted positives that are correctly predicted.

All models were tested under identical conditions, using the same dataset and the same evaluation procedures. This enables a consistent comparison of their performance across these metrics.

7.2.2 Learning curve of traditional Unet

Fig. 7 (a) and (b) display the learning curves for IoU and Cross Entropy loss for training the traditional Unet model with our dataset. The increasing IoU values and decreasing loss in these figures indicate that the model's performance improved steadily during training. After convergence, the model achieved an average IoU score of 0.842 and a loss of 0.00567.

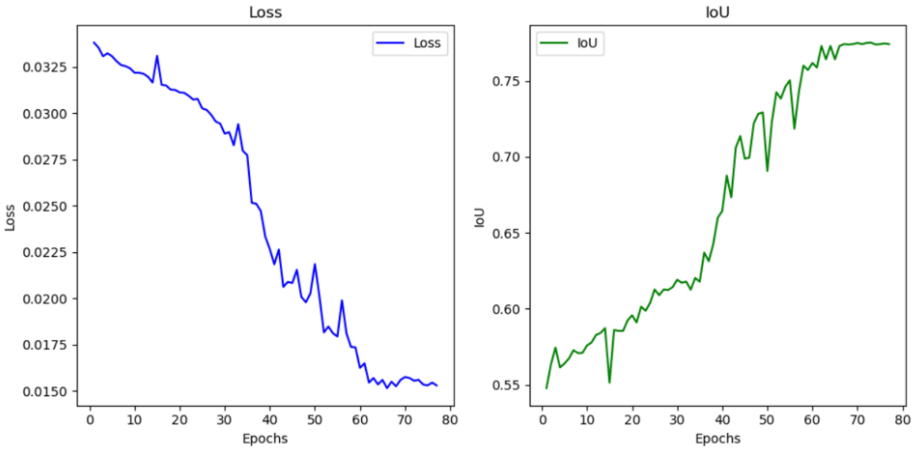

Figure 7. Performance of the baseline UNet model in unclassified cell segmentation, showing IoU and loss metrics.

7.2.3 Learning curve of Self-Attention Unet

We initialized our model using weights pre-trained on the traditional U-Net architecture and further refined it using the same dataset. The progression of training and validation losses is depicted in Fig. 8. The training loss decreased to 1.1×10^{-4}, while the validation loss remained slightly higher, indicating a consistent but slightly conservative generalization.

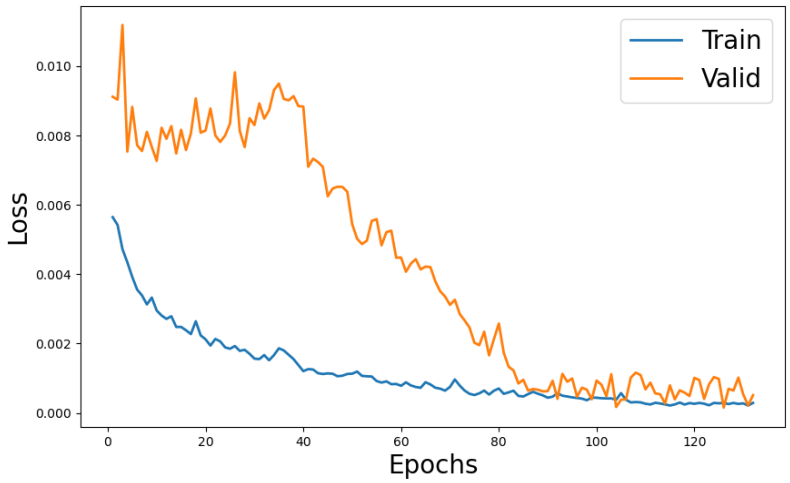

Figure 8. Learning curve for Self-Attention Unet on 10×10 magnification cell images

7.2.4 Confusion matrix on 40× magnification cell images

We also applied our trained model to 562 SLIM images not used in training. We compared the dominant semantic label between the predicted cell nuclei and the ground truth for individual nuclei. The result is shown in Tab. 3.

Through Tab. 3, we could easily distinguish live and dead cells with an error rate of less than 1%. The model achieved high precision, recall, and F1 scores for both live and dead cells, with all metrics exceeding 96%. These results highlight the model's reliability in accurately segmenting and classifying live and dead CHO cells from 40× magnification images, ensuring robust performance in challenging microscopy data.

7.2.5 Velocity Correlations and Displacement Moments of model

We validated the performance of our model by comparing the spatial and temporal velocity correlations, as well as the displacement moments, between the predicted results and the ground truth data. The results are shown in Fig. 9; the spatial

		Ground truth	
		Live ($n = 101718$)	Dead ($n = 99083$)
Cells	Live	99.2%	0.8%
	Dead	0.8%	99.2%
Evaluation	Precision	96.4%	97.5%
	Recall	96.6%	97.9%
	F1 Score	96.5%	97.7%

Table 3. Pixel-wise model prediction and evaluation metrics for live and dead CHO cell images with 40× magnification. Confusion matrix entries are normalized relative to the number of cells in each class.

velocity correlation of our model's prediction exhibits strong agreement with the ground truth across spatial, with minor deviations observed at larger distance scales (Fig. 9a), validating the model's ability to capture the spatial dependency of cell movements. The trends of temporal velocity correlations (Fig. 9b) observed demonstrate high consistency between different temporal scales, capturing the temporal coherence of cell movements. For the scaling behaviors of displacement moments(Fig. 9c), the predicted moments align well with the ground truth, capturing both linear and nonlinear scaling behaviors. These results validate the model's ability to predict cell movement dynamics with high accuracy.

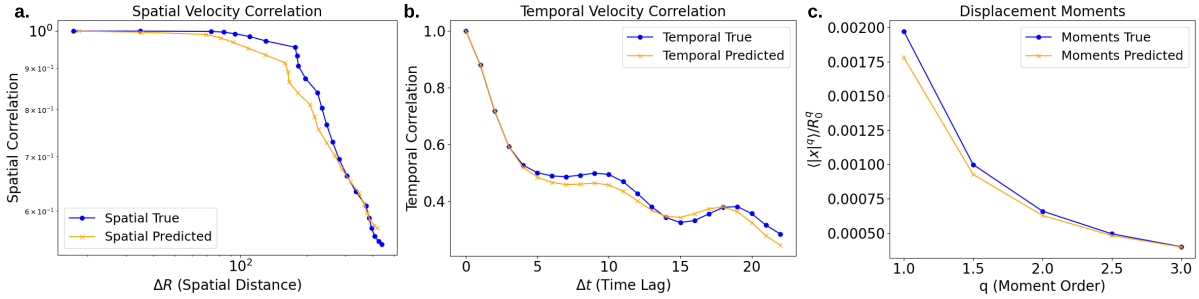

Figure 9. Validation of predicted spatial and temporal velocity correlations and displacement moments. a. Spatial velocity correlation as a function of spatial distance (ΔR) for both the predicted and ground truth data. **b.** Temporal velocity correlation as a function of temporal lag (Δt) comparing predicted and ground truth data. **c.** Displacement moments normalized by the maximum displacement (R_0) for moment orders $q = 1.0, 1.5, 2.0, 2.5, 3.0$.

7.2.6 Model Complexity Comparision

To assess the efficiency and performance of our model, we compared it against Unet and a transformer-based segmentation model (ViT-B-16). The results, summarized in Tab. 4, indicate that despite having 66.34 million trainable parameters, our model maintains a significantly lower computational burden than ViT-B-16 (88.57M) while outperforming U-Net (31.55M) in accuracy. Additionally, our model exhibits a 40.5% reduction in execution time compared to ViT-B-16, requiring 75.43 ms per training versus 135.60 ms for the transformer-based model.

Memory consumption is a critical factor in real-world deployment. While ViT-B-16 requires 12.05 GB of RAM, our model reduces this demand to 10.85 GB, improving efficiency without sacrificing performance. Compared to U-Net, which has the lowest memory footprint (3.59 GB) and execution time (30.36 ms), our model provides a balance between computational efficiency and accuracy.

	Unet	ViT-B-16	Our model
RAM usage (GB) ↓	3.59	12.05	10.85
Execution time (ms) ↓	30.36	135.60	75.43
Trainable parameters(M) ↓	31.55	88.57	66.34
Accuracy(%) ↑	84.2	84.7	**85.8**

Table 4. Comparison of segmentation models in terms of computational efficiency and accuracy. The table presents a comparative analysis of U-Net, ViT-B-16(Transformer), and our proposed model in terms of RAM usage, execution time, number of trainable parameters, and accuracy. All computational metrics are evaluated based on one batch of training.