Computers in Biology and Medicine 182 (2024) 109151

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Accurate detection and instance segmentation of unstained living adherent cells in differential interference contrast images

Fei Pan [a,b,1], Yutong Wu [c,1], Kangning Cui [b,c], Shuxun Chen [d], Yanfang Li [d,e], Yaofang Liu [b,c], Adnan Shakoor [f], Han Zhao [d], Beijia Lu [c], Shaohua Zhi [a], Raymond Hon-Fu Chan [b,c,g], Dong Sun [d,*]

[a] School of Interdisciplinary Studies, Lingnan University, Lau Chung Him Building, 8 Castle Peak Rd — Lingnan, Tuen Mun, New Territories, Hong Kong Special Administrative Region, China
[b] Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE), Room 1115–1119, Building 19 W, Hong Kong Science Park, Hong Kong Special Administrative Region, China
[c] Department of Mathematics, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong Special Administrative Region, China
[d] Department of Biomedical Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong Special Administrative Region, China
[e] School of Communication Engineering, Hangzhou Dianzi University, Qiantang District, Hangzhou, Zhejiang Province, China
[f] Control and Instrumentation Department, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[g] School of Data Science, Lingnan University, 8 Castle Peak Rd — Lingnan, Tuen Mun, New Territories, Hong Kong Special Administrative Region, China

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.13120679

Keywords:
Adherent cell
Cell detection
Cell instance segmentation
DIC images

ABSTRACT

Detecting and segmenting unstained living adherent cells in differential interference contrast (DIC) images is crucial in biomedical research, such as cell microinjection, cell tracking, cell activity characterization, and revealing cell phenotypic transition dynamics. We present a robust approach, starting with dataset transformation. We curated 520 pairs of DIC images, containing 12,198 HepG2 cells, with ground truth annotations. The original dataset was randomly split into training, validation, and test sets. Rotations were applied to images in the training set, creating an interim "α set." Similar transformations formed "β" and "γ sets" for validation and test data. The α set trained a Mask R-CNN, while the β set produced predictions, subsequently filtered and categorized. A residual network (ResNet) classifier determined mask retention. The γ set underwent iterative processing, yielding final segmentation. Our method achieved a weighted average of 0.567 in $AP^{bbox}_{0.75}$ and 0.673 in $AP^{segm}_{0.75}$, both outperforming major algorithms for cell detection and segmentation. Visualization also revealed that our method excels in practicality, accurately capturing nearly every cell, a marked improvement over alternatives.

1. Introduction

The cell, as the fundamental unit of life, exhibits a complex system of material metabolism, energy conversion, and information regulation. In a typical bacterial or animal cell, approximately 70% of its weight consists of water, rendering it transparent and colorless [1].

Most of the cells derived from vertebrates, such as birds and mammals, are adherent cells, excluding hematopoietic cells, germ cells, and a few others. Adherent cells differ from suspension cells as they rely on anchorage to a tissue culture-treated substrate for adhesion and spreading, as depicted by scanning electron microscope (SEM) images in Fig. 1. The irregular morphology of adherent cells contrasts with the spherical shape of suspension cells, presenting challenges for algorithms in detection, segmentation, tracking, and analysis [2,3].

Furthermore, adherent cells' transparency makes them nearly invisible under a light microscope without staining. Hence, researchers commonly use differential interference contrast (DIC) microscopes, capable of observing delicate structures in living or unstained samples and generating 3D images. The working principle involves converting the phase difference of an object into amplitude changes through coherent light beam interference, within a distance of just 1 μm or less, inside and outside the sample.

* Corresponding author.
E-mail addresses: fei.pan@ln.edu.hk, fpan@hkcoche.org, fei.pan@my.cityu.edu.hk (F. Pan), yutwu3-c@my.cityu.edu.hk (Y. Wu), kangnicui2-c@my.cityu.edu.hk (K. Cui), shuxuchen2@cityu.edu.hk (S. Chen), yanfangli2-c@my.cityu.edu.hk, yanfangli@hdu.edu.cn (Y. Li), yaofanliu2-c@my.cityu.edu.hk (Y. Liu), ashakoor2@um.cityu.edu.hk (A. Shakoor), hazhao3-c@my.cityu.edu.hk (H. Zhao), beijialu2-c@my.cityu.edu.hk (B. Lu), shaohua.zhi@ln.edu.hk (S. Zhi), raymond.chan@LN.edu.hk, raymond.chan@cityu.edu.hk (R.H.-F. Chan), medsun@cityu.edu.hk (D. Sun).
[1] Fei Pan and Yutong Wu contributed equally to this work.

https://doi.org/10.1016/j.compbiomed.2024.109151
Received 4 October 2023; Received in revised form 4 September 2024; Accepted 10 September 2024
Available online 26 September 2024

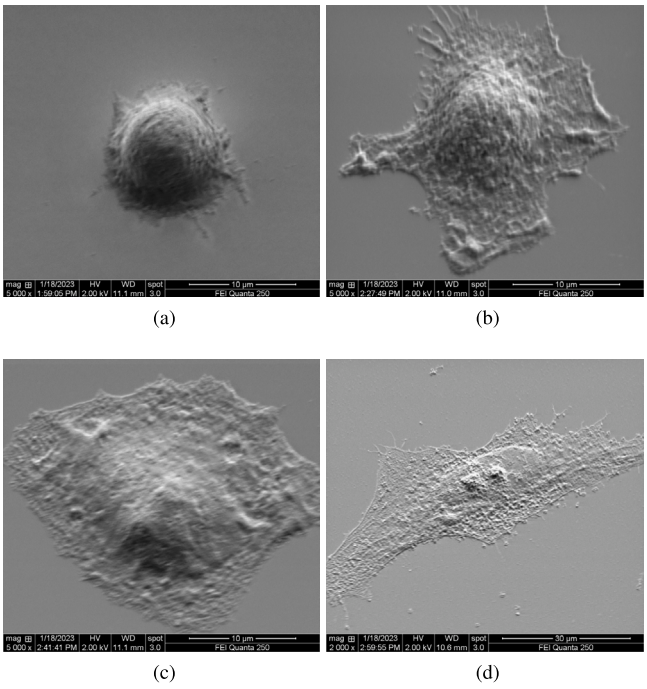


Fig. 1. Morphology of HepG2 cells under an SEM during the cell-spreading process on glass coverslips captured in four steps. Cells were fixed and photographed after (a) 30 min, (b) 60 min, (c) 2 h, and (d) 24 h of attachment.

Fluorescence microscopy is common for observing macromolecules in cells [4]. In this technique, short-wavelength excitation light passes through an excitation filter, causing marked fluorescent molecules to emit visible light. However, fluorescence microscopy has disadvantages like photobleaching and photo-toxicity [5]. So, label-free microscopy is the preferred noninvasive approach for observing living cells [3].

Common tasks of cell image processing include image classification, cell detection/segmentation, tracking, and augmented microscopy [2]. Cell detection locates cell positions using bounding boxes (bboxes), while instance segmentation detects each instance of different cells and generates masks, even if in the same category [6]. Among these tasks, accurate instance segmentation of unstained living adherent cells in DIC images is a common problem in many biomedical experiments, such as cell microinjection [7,8], cell tracking [9], cell activity characterization [10], and revealing cell phenotypic transition dynamics [11]. Solutions to this problem still need improvement, though several data-independent and machine/deep-learning algorithms have been proposed to deal with the challenge of instance segmentation for various cell lines under different imaging modalities [3,12].

The difficulty of instance segmentation for unstained adherent cells arises from at least four main challenges. Firstly, certain elongated cells are both inclined and closely fused, making a region proposal network (RPN) challenging to propose separate bboxes to differentiate them as distinct instances. Consequently, such cells often remain undetected by the RPN, resulting in their non-segmentation. Secondly, an RPN can propose separate bboxes encompassing certain cells but the bboxes overlap remarkably, leading to suboptimal mask segmentation. Such cells tend to be treated as one large cell by the non-maximum suppression (NMS) after bbox regression, causing them to appear merged in the final segmentation outputs. Third, cells manifest diverse health statuses, encompassing both healthy and unhealthy states. Yet, many datasets and algorithms predominantly concentrate on cells in a singular healthy state. Last but not least, there was little high-quality DIC cell image dataset. Currently, larger-scale datasets primarily consist of phase contrast (PhC) images and fluorescence images, such as the cell tracking challenge (CTC) dataset [9], the LIVECell dataset [13], and the TissueNet dataset [14].

In light of these considerations, our present study aims to contribute two key elements: (1) the creation of a new, high-quality dataset comprising finely annotated, high-resolution DIC images of unstained living adherent cells; and (2) the development of a novel cascaded method for detection and instance segmentation of these cells.

Our dataset comprises 520 DIC images of 12,198 unstained HepG2 human liver cancer cells, each with a corresponding fluorescence image stained with calcein acetoxymethyl (AM), ensuring high-quality ground-truth annotations. Unique in addressing the multi-state nature of adherent cells commonly seen in wet labs, it includes both healthy and unhealthy cells in a single image, providing a valuable resource for studying multi-state cell detection and instance segmentation.

Our method is derived from an intuitive question: can we rotate all input images by a certain degree and use both the original and rotated images as inputs to improve bbox predictions? By doing so, appropriate bboxes can be generated for easily overlooked and confused cells.

Before proceeding to our method, we randomly divide our dataset into three distinct subsets: a training set, a validation set, and a test set. Subsequently, we performed 45° rotations on the images and their corresponding annotations within the training set. This process yielded an interim dataset, denoted as the "α set", comprised of both the original and rotated images, alongside their annotations. This rotated image and the original image together form a dual-view input (DVI), which offers several advantages: it enhances the coverage of bboxes for those overlooked cells without considerably increasing types or counts of predefined bboxes and concurrently diminishes bbox overlap for those confused cells.

Similar procedures were applied to the validation and test sets, resulting in the creation of the "β set" and "γ set", respectively. The α set was used to train a Mask R-CNN [15] with relaxed NMS. Following training, the β set underwent processing through the Mask R-CNN, generating a substantial number of predictions. These comprehensive yet redundant predictions were subjected to a connectivity-based filtering and categorization process. Predictions featuring the highest intersection over union (IoU) of masks are retained and categorized as "image (mask) patches to retain", whereas the remaining predictions are classified as "image (mask) patches to discard". This categorization bifurcates the image (mask) patches into two distinct groups.

A residual network (ResNet) classifier [16] was then trained on these patches, determining which masks were to be retained. The use of ResNet is referred to as supervised mask selection (SMS) because it avoids unsupervised NMS of bboxes on the initial predictions from Mask R-CNN and can efficiently classify numerous redundant outputs. In the final step, the γ set underwent iterative processes mirroring the previous steps, culminating in the derivation of the ultimate instance segmentation results.

Our method outperforms major object detection and instance segmentation algorithms on our dataset, achieving a weighted average of 0.567 in $AP^{bbox}_{0.75}$ and 0.673 in $AP^{segm}_{0.75}$, accurately capturing nearly every cell. In summation, our method optimizes instance segmentation for adherent cells, mitigating challenges posed by overlapping and inclined cells and accommodating various cell health states. This study will directly benefit direct image cytometry by providing consistent, quantitative, and informative cell measures and revealing cellular heterogeneity not easily detectable by humans. It will also aid various downstream analyses and experiments relying on accurate cell instance segmentation, such as cell activity analysis, cell phenotyping, cell tracking, cell microinjection, and digital histopathology.

The article is organized as follows: Section 2 introduces instance segmentation of generic objects and adherent cells. Section 3 describes our dataset, while Section 4 provides details of our method. Section 5 reports quantitative and qualitative comparisons with other algorithms, and Section 6 discusses dataset and algorithm limitations. Finally, Section 7 concludes the study.

2. Related work

Instance segmentation has become a vital pursuit in computer vision, especially in fields like biomedical imaging. It aims to precisely identify and segment individual instances of various object categories within an image. Deep learning approaches for instance segmentation generally fall into two categories: one-stage and two-stage methods [17,18]. One-stage methods usually use convolutional neural network (CNN) models directly for instance segmentation, omitting explicit feature localization. Conversely, two-stage methods, often based on R-CNN, follow a two-step process: first, detecting bboxes containing objects, and then predicting foreground—background masks for each region of interest (RoI), as exemplified by Mask R-CNN.

Mask R-CNN, a milestone in generic instance segmentation, extends Faster R-CNN [19] by integrating a mask branch for precise instance delineation. Like Faster R-CNN, it employs an RPN to identify RoIs in each input image. Subsequently, the model predicts the category and spatial layout of the contained object.

Mask Scoring R-CNN [20] marks an advancement in instance segmentation. It addresses a common challenge where the quality of predicted masks does not always align with the classification score. To tackle this, Mask Scoring R-CNN introduces a specialized network block that learns and fine-tunes mask quality, resulting in improved performance in instance segmentation.

In situations involving densely clustered or irregularly shaped objects, Rotated Mask R-CNN [21] introduces a novel approach. It effectively overcomes a limitation in Mask R-CNN, which struggles with scenes containing multiple objects of the same class with high bbox overlap. Unlike conventional methods that assume a single object per bbox, Rotated Mask R-CNN employs rotated bboxes. This innovative representation improves segmentation accuracy, particularly in real-world applications like robotics, logistics, and household objects, where such scenarios are common but often underrepresented in standard datasets.

Similar to Rotated Mask R-CNN, Oriented R-CNN for object detection [22] and the subsequent Oriented R-CNN for instance segmentation [23] were introduced in 2021 and 2024, respectively. Oriented R-CNN is mainly applied to object detection in remote sensing images, where the oriented RPN generates arbitrarily oriented bbox proposals and uses a lightweight fully convolutional network to improve efficiency. Specifically, it achieves this by increasing the RPN regression branch output parameters from four to six, using a midpoint offset representation scheme.

Recently, PointRend [24] introduced a novel approach to instance segmentation. Unlike traditional methods, it fine-tunes mask predictions at specific points within the mask rather than uniformly. This targeted refinement results in more precise object boundaries, mitigating the issue of excessive smoothing seen in earlier techniques.

QueryInst [25] introduced a novel approach to instance segmentation, presenting a multi-stage end-to-end system that treats instances as adaptable queries. Attributes such as categories, bboxes, instance masks, and instance association embeddings are unified under this query framework. QueryInst uniquely shares a query for both detection and segmentation, achieved through dynamic convolutions and parallelly-supervised multi-stage learning.

In recent developments, attention-based transformers called DETR or DEtection TRansformer [26] have shown promise in object detection and universal image segmentation. It introduces a pioneering approach by treating object detection as a direct set prediction problem, bypassing the need for hand-designed components like NMS and bbox generation. It leverages a set-based global loss and a transformer encoder–decoder architecture to ensure precise predictions through bipartite matching. Furthermore, DETR uses a fixed set of learned object queries to reason about object relations and global image context, streamlining the detection process.

While prior reviews [2,3,12] have covered instance segmentation of adherent cells, some algorithms warrant attention. StarDist [27], developed in 2018, was tailored for instance segmentation of stained nucleus images [28]. Unlike traditional methods that rely on bboxes or pixel grouping, StarDist employs star-convex polygons as a more accurate representation of cell shapes. This eliminates the need for subsequent shape refinement and addresses segmentation errors in crowded cell scenarios.

In 2019, ANCIS [29] was proposed to detect and segment individual rat neural stem cells (NSCs) in DIC images. By combining a single-shot multi-box detector (SSD) [30] and a U-Net [31] within a unified network, ANCIS simultaneously predicts precise bboxes and segmentation masks for each cell. This method leverages attention mechanisms in both the detection and segmentation modules to focus on critical features.

Also in 2019, a CNN-watershed-mixed algorithm was introduced to detect and segment individual cells in PhC and DIC images through a three-step process [32]. The initial step involved training a CNN to learn the Euclidean distance transform of an input mask, followed by training a Faster R-CNN to detect individual cells in the output image from the first step. Finally, a watershed algorithm was applied to the outputs of the preceding two steps to produce the final segmentation.

In 2020, Cell-DETR [33], inspired by DETR, was introduced for detecting and segmenting cells in microstructures. It excels in segmenting yeast in microstructured environments, surpassing existing methods for semantic segmentation and offering individual object instance predictions. However, it downsizes images before feature extraction, which may potentially impact its ability to capture fine details.

In the same year, a test-time augmentation (TTA)-enhanced Mask R-CNN [34] was developed. The proposed method rotates input images by 90°, 180°, and 270° at the inference stage and uses a combination of "object matching" and "majority voting" strategy[2] to decide whether to keep a prediction. If one predicted object appears in most of the binary mask images, it will be kept by voting. However if another predicted object appears in only a few binary mask images, it will be excluded by voting.

In 2021, a novel weakly supervised algorithm, which relies solely on approximate cell centroid positions for training, was developed for cell segmentation [35]. This algorithm comprises three steps: firstly, a cell-detection CNN was trained using rough cell centroid positions. Second, region back-propagation was employed to extract a cell relevance map. Finally, graph-cut was applied to the relevance map obtained in the second step to generate the final instance segmentation. This algorithm reduces the annotation burden compared to standard supervised methods.

In the same year, Cellpose [36] was developed for cellular segmentation as a generalist algorithm. The authors of Cellpose found that classical segmentation approaches based on the watershed algorithm did not work well on fluorescence images of stained cells. These images often have issues like the nuclear exclusion of the fluorescent marker and its uneven distribution along cell borders and protuberances, leading to multiple intensity basins. To address this, they aimed to construct an intermediate representation of an object that forms a single smooth topological basin.

Their method consists of three steps. First, they generated topological maps from fluorescence images of stained cells through simulated diffusion using ground-truth masks drawn by a human annotator. Second, they trained a modified U-net on these topological map datasets to predict the horizontal and vertical gradients of the maps and a binary map indicating if a pixel is inside or outside RoI. Third, they used gradient tracking on the U-net predictions to route all pixels belonging to a cell to its center, grouping them to recover individual cell shapes.

[2] Majority voting: An object must be detected in the majority of the images (original and augmented) to be included as the final mask.

TTA was used to improve predictions. Although Cellpose was tested on various cell images, it currently lacks direct support for simultaneous multi-class instance segmentation.[3]

More recently, a novel box-based instance segmentation network was created for detecting and segmenting leaf and cell images [37]. It uses an object-guided strategy to distinguish targets from neighboring objects with similar textures and low-contrast boundaries. It separates target objects from adjacent ones within the same bbox region by first detecting object center points and predicting bbox parameters, followed by object-guided coarse-to-fine segmentation branches. Additionally, an auxiliary feature refinement module is introduced to enhance segmentation quality by densely sampling and refining point-wise features in boundary regions.

3. Dataset

3.1. Dataset curation

This study utilized the cell micromanipulation platform developed by the authors previously [7] to automatically collect images of *in vitro* cultured HepG2 human liver cancer cells (grown in a Petri dish). Our hardware platform mainly consists of an inverted fluorescence microscope (Eclipse Ts2R-FL, Nikon) with a 40× objective lens (CFI S Plan Fluor ELWD 40XC 228 MRH08430, Nikon), a motorized XY stage (ProScan H117P1N4, Prior Scientific), and a complementary metal oxide semiconductor (CMOS) camera (DigiRetina 16, Tucsen Photonics). With this hardware configuration, all raw red, green, and blue (RGB) images captured by the camera were downscaled from 2304 pixel × 1728 pixel to 1152 pixel × 864 pixel using OpenCV's default bilinear interpolation algorithm, representing approximately 216.500 μm × 162.375 μm in the dish.

The curation of the cell image dataset involves two main steps: image acquisition and post-annotation.

The image acquisition process includes several steps. First, cells were cultured in Dulbecco's modified eagle medium (DMEM) (Gibco) supplemented with 10% fetal bovine serum (FBS) (Gibco), 100 U/mL of penicillin, and 100 U/mL of streptomycin in a 35 mm glass-bottom petri dish (culture dish 801002, Wuxi NEST Biotechnology) and placed in a humidified atmosphere of 37 °C and 5% CO_2. Second, Calcein AM, a commonly used fluorescent dye, was used to test cell viability and for short-term staining. Living cells stained with calcein AM emit green fluorescence at a wavelength of 530 nm when excited by 488 nm light (fluorescence mode of the microscope), whereas dead cells do not stain and therefore are not visible in fluorescence mode. According to the dye's instructions, 5 μL of 4 mmol calcein AM (L6037S, US Everbright, Inc.) was taken from the refrigerator and brought to room temperature before image collection. It was then mixed with 10 mL of phosphate-buffered saline (PBS) to stain the cultured cells. Third, the stained cells were taken out of the incubator and placed on the motorized stage. The microscope was set to DIC imaging mode, and the stage was controlled to move the dish in a zigzag pattern while the camera took images at each step. In this study, 2260 DIC images were collected in one session. Fourth, the microscope was switched to fluorescence mode, the stage returned to the starting point, and the same path was repeated to capture corresponding fluorescence images.

The advantage of this image acquisition method is that it ensures the cells are alive under the microscope. In the microscope, they appear different from dehydrated and fixed in vitro cultured cells.

After obtaining the 2260 DIC images and 2260 fluorescence images, we randomly selected 520 pairs of images (DIC and fluorescence), which were then randomly divided into training, validation, and test sets. Since these DIC and fluorescence images correspond one-to-one (e.g., DIC images [Figs. 2(a-1) to 2(c-1)] and their fluorescence counterparts [Figs. 2(a-2) to 2(c-2)]), and the individual living cells in the fluorescence images are easily recognizable (even by inexperienced personnel, as it simply involves outlining the green fluorescent cells in the images), we merged the DIC and fluorescence images (e.g., [Figs. 2(a-3) to 2(c-3)]) and sent them to a professional annotation company. They used an internally modified version of the labelme software [38]. Finally, all images were meticulously checked and corrected by us, with the annotated ground truth saved in a single JavaScript object notation (JSON) file.

Table 1
Statistics of our dataset.

	Train	Val	Test	Total
Avg. healthy cell counts per image	22.5	21.23	19.31	21.61
Avg. unhealthy cell counts per image	1.56	1.96	2.61	1.85
Avg. all cell counts per image	24.06	23.19	21.91	23.46
Avg. healthy cell sizes (μm^2)	747.52	745.82	778.32	752.69
Avg. unhealthy cell sizes (μm^2)	622.37	737.05	648.98	654.18
Avg. all cell sizes (μm^2)	739.4	745.08	762.94	744.92
Avg. healthy cell aspect ratios	1.25	1.21	1.29	1.25
Avg. unhealthy cell aspect ratios	1.25	1.34	1.24	1.27
Avg. all cell aspect ratios	1.25	1.22	1.28	1.25
Percentage of healthy cells	93.51%	91.54%	88.11%	91.05%
Percentage of unhealthy cells	6.49%	8.46%	11.89%	8.95%
Max. objects per image	83	67	53	83

3.2. Dataset characteristics

Adherent cells can be roughly classified into two types from the perspective of cell health: healthy (living) and unhealthy (dead or loosely attached) cells, as depicted in Fig. 2(c-3). Healthy adherent cells typically adhere to the culture surface, have an irregular morphology, and appear entirely green in the fluorescence images once stained with calcein AM. In comparison, some unhealthy cells, e.g., dead cells, can hardly be stained by calcein AM and only appear slightly green. Other unhealthy cells can be successfully stained by calcein AM but loosely adhere to the culture surface and are not ideal candidates for typical biomedical experiments, such as cell microinjection.

In addition to classifying cells by how healthy they are, they can be classified by how densely they grow, as shown in Figs. 2(a-1) and 2(b-1). Sparsely distributed cells [Fig. 2(a-1)] are relatively easy to recognize, but densely distributed cells [Fig. 2(b-1)] are difficult to distinguish from one another, even by humans. Therefore, people can distinguish individual cells clearly [Figs. 2(b-3) and 2(b-5)] only through living cell staining [Fig. 2(b-2)].

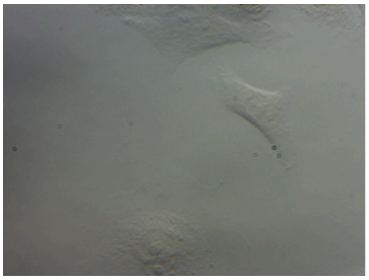
Table 1 presents the basic characteristics of our dataset. It reveals a predominance of healthy cells, indicating a potential need to address category imbalance during algorithm development. Images obtained from multiple experiments exhibit similar patterns. Our images are derived from cells cultivated according to standard procedures, hence the ratio of healthy to unhealthy cells aligns with the expected outcomes of real experiments, albeit somewhat exaggerated. Variability in cell sizes and aspect ratios necessitates adjusting the segmentation approach to accurately capture a wide range of cell shapes. Additionally, the maximum number of objects per image varies across subsets, highlighting the necessity for robust handling of different cell densities during instance segmentation.

4. Methods

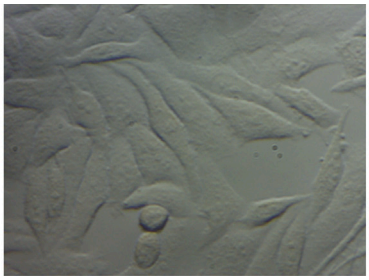
4.1. Introduction to our cascaded instance segmentation method

Our cascaded method is conceptually straightforward, divided into three main steps, as shown in Fig. 3. The first main step consists of two sub-steps. The first sub-step involves randomly dividing our dataset into

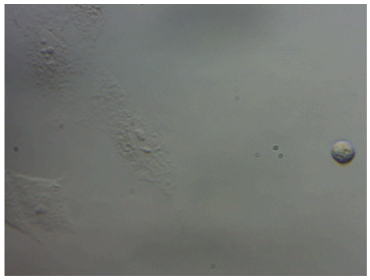
[3] https://cellpose.readthedocs.io/en/latest/faq.html.



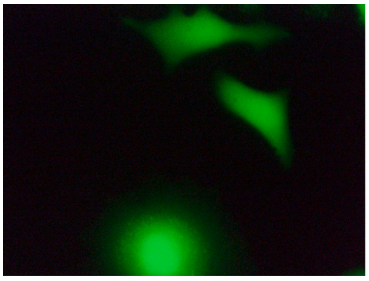
(a-1) A DIC image of sparsely distributed adherent cells.



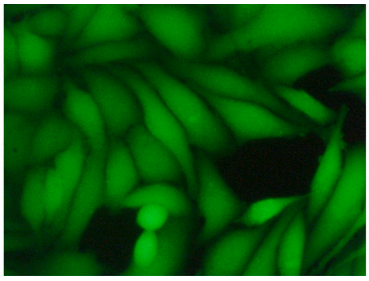
(b-1) A DIC image of densely distributed adherent cells.



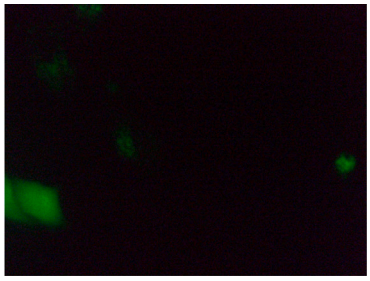
(c-1) A DIC image of unhealthy adherent cells.



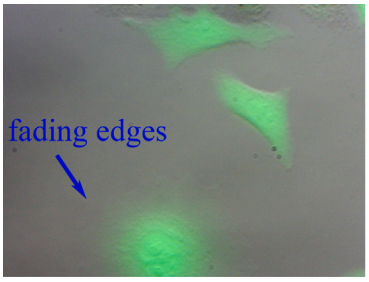
(a-2) A fluorescence image of stained cells in (a-1).



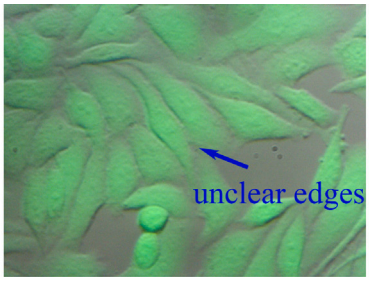
(b-2) A fluorescence image of stained cells in (b-1).



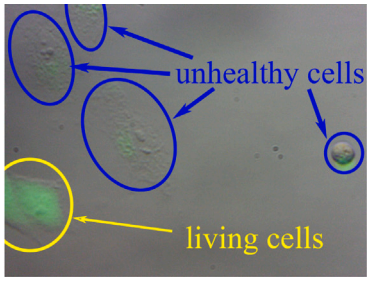
(c-2) A fluorescence image of stained cells in (c-1).



(a-3) A merged image of (a-1) and (a-2).



(b-3) A merged image of (b-1) and (b-2).

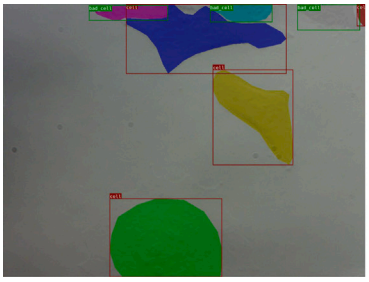


(c-3) A merged image of (c-1) and (c-2).

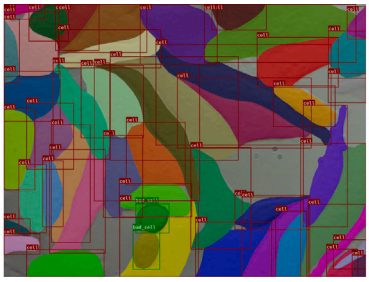
(a-4) Annotated cell health states of (a-1).

(b-4) Annotated cell health states of (b-1).

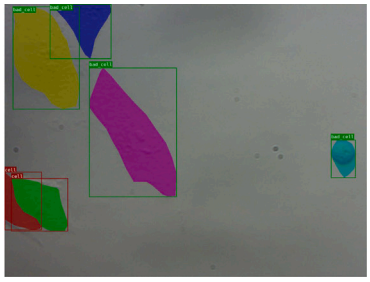
(c-4) Annotated cell health states of (c-1).



(a-5) Annotated ground truth of (a-1).



(b-5) Annotated ground truth of (b-1).



(c-5) Annotated ground truth of (c-1).

Fig. 2. Representative images from our dataset. (a-1) to (c-1) display raw DIC images used for training or interference. (a-2) to (c-2) present fluorescence images. Calcein AM-stained living cells fluoresce green, while unhealthy or dead cells do not. (a-3) to (c-3) are merged images of raw DIC and fluorescence for facilitating manual annotation. (a-4) to (c-4) showcase the annotated cell health states. Finally, (a-5) to (c-5) reveal the annotated ground truth.

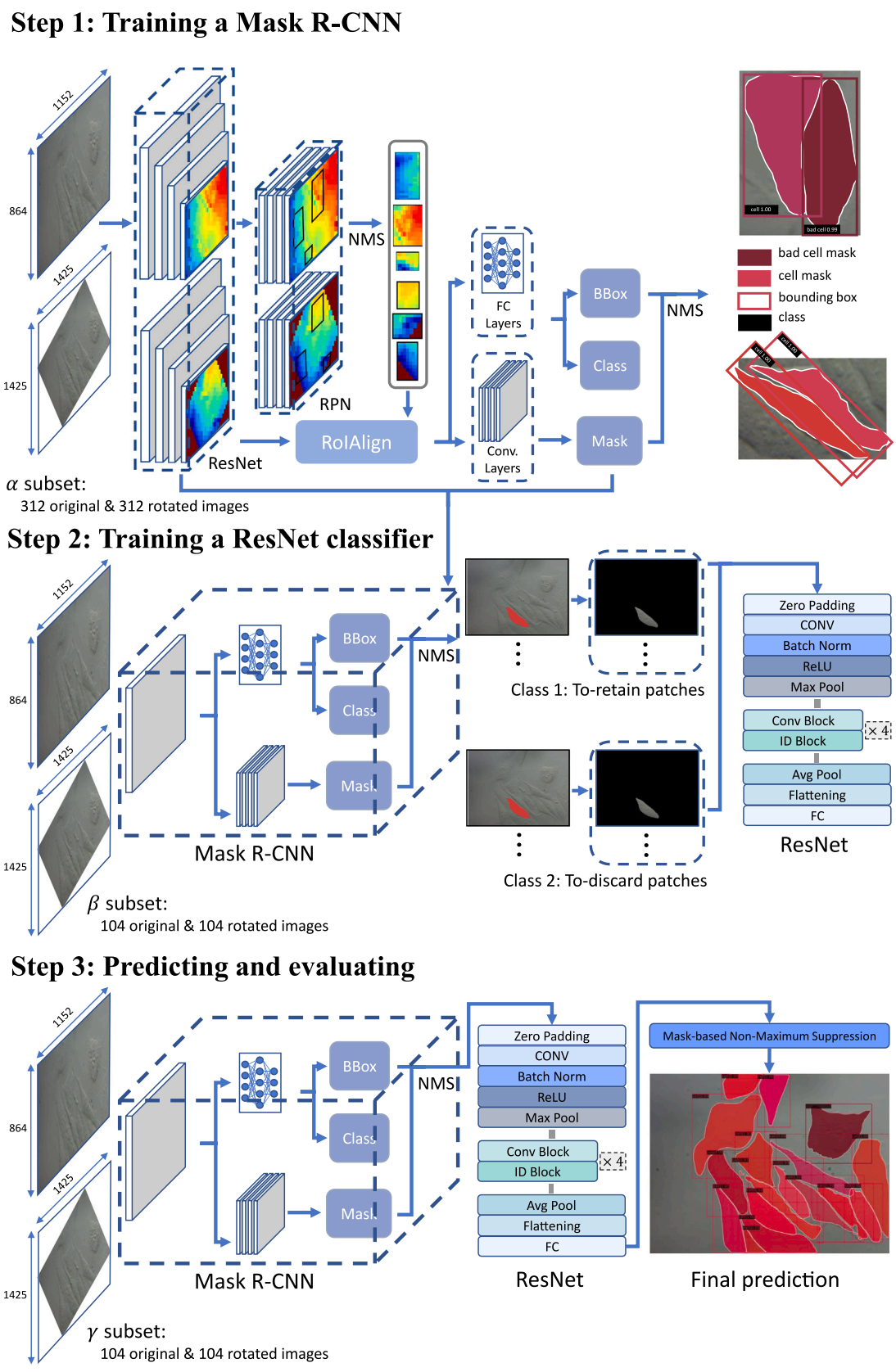
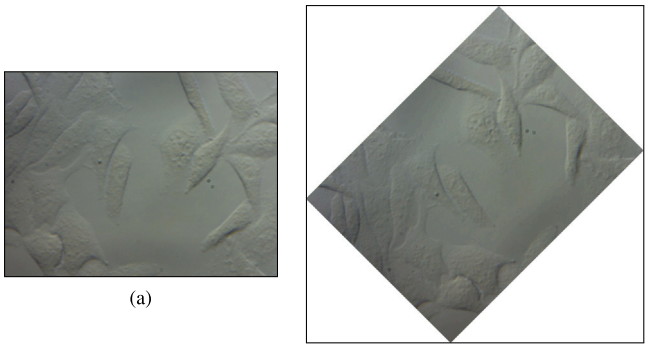


Fig. 3. Overview of our cascaded method. In Step 1, original and rotated images are fused to form the DVI dataset, offering advantages in cell detection. In Step 2, predictions are refined through SMS based on connectivity. Step 3 involves iterative refinement and post-processing, culminating in the final visualized results.



(a) (b)

Fig. 4. An original image and its rotated counterpart.

(a) (b)

Fig. 5. An original mask and its rotated counterpart.

three distinct subsets: a training set, a validation set, and a test set. Then, every image from each subset is rotated by 45° and combined with its original counterpart, forming three DVI subsets, namely "α", "β", and "γ" sets. The second sub-step involves training a Mask R-CNN model on the α subset, incorporating a relaxed NMS for RPN proposal to generate numerous predictions.

Notably, in the first sub-step, we primarily use OpenCV's functions for image and mask rotation. For images, we add white pixels around the rotated image, also referred to as padding with (255, 255, 255) color, as demonstrated in Fig. 4. As for masks, since the background is typically black, we append black pixels around the rotated mask, known as padding with (0, 0, 0) color, as depicted in Fig. 5. The DVI is processed as two separate images rather than being concatenated. We use Python dictionaries and lists to maintain the correspondence between an original image and its rotated counterpart, as well as the predictions (binary masks) for each.

We proposed the DVI technique because when training and predicting with a native Mask R-CNN on both unrotated and rotated images separately, we observed that certain cells were not well detected and segmented in the unrotated images, as indicated by the bboxes in Fig. 6(b) bottom right corner, whereas these cells were adequately detected and segmented at corresponding positions in the rotated images. Similarly, some cells were not well detected and segmented in the rotated images, as seen in the colored bboxes on the left side of Fig. 6(c), but were well-detected and segmented in the unrotated images, as observed in Fig. 6(b).

Therefore, DVI provides two advantages. Firstly, it enhances the visibility of certain cells with slender dimensions, inclined orientation, or close adjacency, which might be overlooked by RPN. Such cells, as delineated in Fig. 7(a-1), become more apparent after rotation, aligning vertically or horizontally, as exemplified in Fig. 7(a-2). Secondly, cells that are susceptible to confusion after bbox regression and NMS, as shown in Fig. 7(b-1), gain greater clarity, as shown in Fig. 7(b-2). DVI has similarities with TTA in this article [34] and offers us a more comprehensive set of output details.

In the second sub-step, the relaxed NMS for the RPN proposal ensures the preservation of all possible predicted masks, including those that might be redundant or overlapping, thus retaining valuable information from both original and rotated perspectives.

In the second main step, the β subset undergoes processing with the trained Mask R-CNN. This processing yields a substantial number of predictions, ranging from those requiring minimal refinement to those demanding further attention. These predictions are subjected to a connectivity-based filter, facilitating the retention of masks characterized by simple connectivity. These retained masks, along with their associated bboxes and classifications, are then compared to the ground truth annotations. This results in a two-fold categorization based on the IoU measurement of masks. Image (mask) patches classified as "to retain" constitute Class 1, while those classified as "to discard" comprise Class 2. The two categories are merged into a provisional dataset for training a ResNet classifier.

This process, termed SMS, effectively replaces the second round of NMS within the Mask R-CNN. In SMS, we employ a two-step approach to refine and select the best segmentation masks. First, we use a trained ResNet classifier to evaluate the predicted masks, classifying them as either valid or invalid. This allows us to filter out low-quality masks. For the valid masks, we apply a comparison process based on connected components and IoU. Masks with more than six connected components are discarded to eliminate noise. For the remaining masks, we compute the IoU between each pair and use a scoring system that combines connected components count, IoU, and classification confidence. The mask with the lower score is discarded, while the higher-scoring one is retained.

SMS ensures that only the most accurate masks are selected by combining classification, IoU, and connected components analysis. As a result, we improve segmentation quality and suppress false positives or redundant masks, providing a refined set of binary masks for further analysis.

SMS proves advantageous in scenarios with dense cell arrangements, where NMS of bbox might fall short because the confidence score of each prediction only reflects the likelihood of an object's presence, without considering the accuracy or proximity of a predicted mask to the shape of an actual object. Hence, a few bboxes with low confidence scores might still predict correct masks.

In the third main step of our method, the γ subset undergoes an iterative process involving Steps 1 and 2. Once the ResNet classifier generates a collection of "to-retain image patches", this set is subjected to another round of SMS to derive the final masks. Finally, masks that have the largest IoU (and also over 0.7) with the other masks at "each spot" are preserved to prevent redundancies, as shown in Fig. 8.

4.2. Evaluation metrics

For our assessment, we employed common objects in context (COCO) metrics due to their standardized application programming interfaces (APIs), which seamlessly support detection and instance segmentation across various object categories within a single image. This standardized approach allows direct comparisons between general-purpose algorithms and those tailored for cell analysis. We used 13 advanced metrics based on fundamental measurements including IoU, precision, and recall to evaluate algorithm performance (see Table 2).

IoU (Jaccard index) quantifies object detection and segmentation accuracy by measuring the overlap between predicted and ground truth regions. Precision identifies correctly detected objects[4] out of

[4] true positives (TPs).

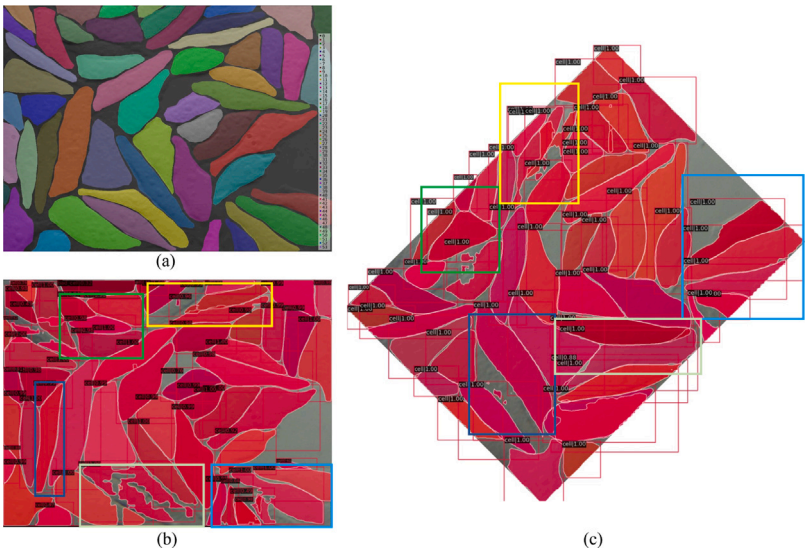


Fig. 6. Illustration of certain cells overlooked by a classic Mask R-CNN. (a) Ground truth of the sample cell image. (b) Inference results of the sample cell image show that the cells detected in the upper bboxes are missed in (c). (c) Inference results of the rotated sample cell image reveal that the cell detected in the lower right bboxes is missed in (b).

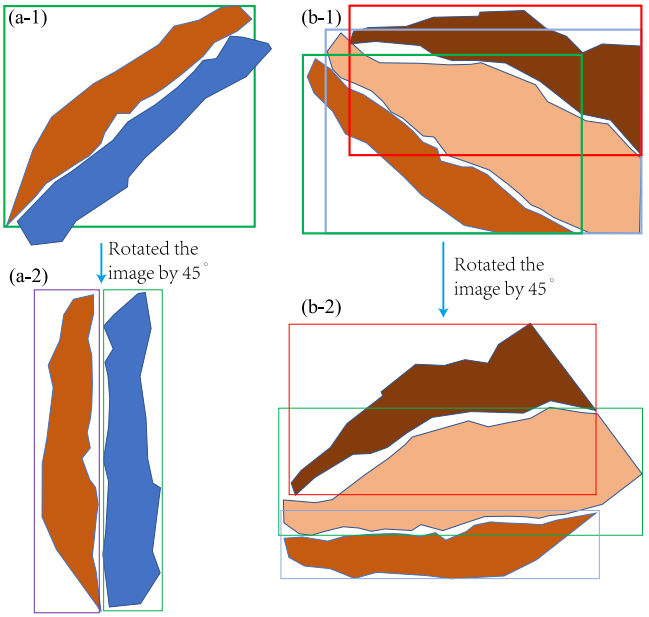


Fig. 7. Illustration of image rotation and its benefits. (a-1) Illustrates certain cells prone to being overlooked by the RPN due to their slender dimensions, inclined orientation, or close adjacency. (a-2) Demonstrates enhanced visibility of these cells after a 45° rotation, rendering them vertically or horizontally aligned and more discernible. (b-1) Depicts cells susceptible to confusion after bbox regression and NMS. (b-2) Highlights the increased clarity of these cells after undergoing the proposed approach.

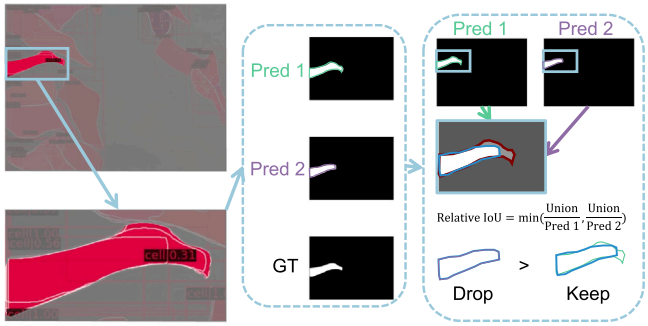


Fig. 8. Mask prediction selection. We finally keep the smaller mask prediction, which tends to contain more comprehensive information.

all predicted objects,[5] while recall measures the fraction of correctly detected objects out of all ground truth objects.[6] Higher precision signifies fewer false positives and higher recall indicates better capture of true positives.

AP represents the mean precision across different recall levels. It is widely used in object detection tasks, calculated by generating the precision–recall curve for a specific category and computing the area under the curve (AUC). AP offers a comprehensive measure of model performance across various categories and serves as a common benchmark for comparing different object detection models. Higher AP or mAP indicates superior object detection across all object categories in the dataset. Notably, AP^{bbox} focuses solely on bbox accuracy, while AP^{segm} incorporates both bbox and pixel-wise segmentation accuracy.

Specifically, AP (at IoU = 0.50 : 0.05 : 0.95) computes average precision at multiple IoU thresholds ranging from 0.50 to 0.95 with an increment of 0.05. This evaluates the model's performance across a wide range of IoU thresholds, capturing its ability to detect objects with varying degrees of overlap between predicted and ground truth regions. AP (at IoU = 0.50) is commonly used for comparison with results from the pattern analysis, statistical modelling, and computational learning (PASCAL) visual object classes (VOC) dataset, which employs this specific IoU threshold for object detection evaluation. AP (at IoU = 0.75) is calculated at a higher IoU threshold of 0.75, assessing the model's capability to produce more accurate bboxes or segmentation masks by requiring a higher level of overlap between predicted and ground truth regions.

Average recall (AR) is computed as the average of TPs divided by the sum of TPs and FNs across all categories. A high AR score indicates effective capture of positive cases without many omissions. Specifically, AR_1 evaluates a model's performance when considering only the most confident detection for each object in the image. AR_{10} considers up to 10 of the most confident detections per object, striking a balance between precision and recall. AR_{100} considers up to 100 of the

[5] TPs + false positives (FPs).
[6] TPs + false negatives (FNs).

most confident detections per object, providing a more comprehensive assessment of the model's localization capabilities.

Despite the above metrics, we also follow this paper [29] and use the average IoU at thresholds α (also known as instance-wise mean IoU) for performance evaluation. It is defined as $\mathrm{AIoU}_\alpha = \frac{1}{N_\alpha}\left(\sum_{i=1}^{N_\alpha}(\mathrm{IoU})_i\right)$, where N_α is the total number of predictions that satisfy IoU $\geq \alpha$.

4.3. Implementation details

Here we denote our original dataset as ϕ, comprising 520 images and associated annotations encoded in a COCO-style JSON file. To introduce diversity and address intricate segmentation scenarios, all images undergo a 45° rotation. This rotation is propagated to align the ground truth annotations with the rotated images. Merging the rotated images and annotations with their unrotated counterparts and original annotations yields an adaptable interim dataset, referred to as Φ. For focused analysis, we partition Φ into three subsets: α, β, and γ. The α subset includes 312 unrotated images, their rotated versions, and the associated annotations. The β subset contains 104 unrotated images, their rotated counterparts, and the corresponding annotations. The γ subset contains the rest 104 unrotated images, rotated versions, and annotations.

5. Experimental results

This section presents a comprehensive comparison of our method with several existing algorithms.

5.1. Quantitative results

Our research aims to achieve simultaneous detection and instance segmentation of adherent cells with different health states within the same image. We selected StarDist, ANCIS, Rotated Mask R-CNN, and four generic instance segmentation algorithms (Mask R-CNN, Mask Scoring R-CNN, PointRend, and QueryInst) from the MMDetection toolbox to compare with our method. Among these algorithms, ANCIS only natively supports the detection and segmentation of a single category of cells in a single image. However, since our task required multi-category segmentation, we revised ANCIS's code to handle both COCO-style and Kaggle-style ground truth annotations (ANCIS' original annotation style), enabling it to perform multi-category segmentation. We trained and tested the adapted ANCIS on the large-scale LIVECell dataset, which contains single-category cell PhC images. The Kaggle-style results ($\mathrm{AP}^{\mathrm{segm}}_{0.50} = 0.264$) and the COCO-style results ($\mathrm{AP}^{\mathrm{segm}}_{0.50} = 0.281$) were very similar, with minor differences in AP calculation. Based on this observation, we believe that our adaptation is effective.

Detailed comparison results are provided in Table 3. Our method achieves a weighted average of 0.567 in $\mathrm{AP}^{\mathrm{bbox}}_{0.75}$ and a weighted average of 0.673 in $\mathrm{AP}^{\mathrm{segm}}_{0.75}$. Both metrics are very stringent because they require a high IoU threshold of 0.75. Our method's performance in delineating object boundaries accurately is well demonstrated, outperforming other methods listed in Table 3. Fig. 9 further illustrates the accuracy of our predictions. These visual results provide clear evidence that our model achieves high quantitative scores under strict metrics and produces visually accurate and reliable predictions in practical scenarios. However, our method still has room for improvement in detecting and segmenting unhealthy cells. While certain metrics may outperform the other algorithms, there remains some gap when compared to our method's AP values for healthy cells.

In addition, we conducted a comparison with Cell-DETR, an attention-based transformer. However, we encountered certain limitations in its native code, such as the lack of support for RGB images and the small resolution (downsized to 128 pixel × 128 pixel) of input and output. Despite setting the query count to 100, the algorithm's built-in metrics showed a mIoU of only around 0.27.

Table 2
Basic and advanced evaluation metrics for cell detection and instance segmentation.

Metrics	Explanation
IoU	$\mathrm{IoU(GT, P)} = \frac{\mathrm{area(GT \cap P)}}{\mathrm{area(GT \cup P)}}$ [a]
Precision	$\mathrm{Precision} = \frac{\sum \mathrm{TP}}{\sum \mathrm{TP}+\sum \mathrm{FP}} = \frac{\sum \mathrm{TP}}{\mathrm{all\ detections}}$
Recall	$\mathrm{Recall} = \frac{\sum \mathrm{TP}}{\sum \mathrm{TP}+\sum \mathrm{FN}} = \frac{\sum \mathrm{TP}}{\mathrm{all\ ground\ truths}}$
Precision(τ)	$\mathrm{Precision}(\tau) = \frac{\sum \mathrm{TP}(\tau)}{\sum \mathrm{TP}(\tau)+\sum \mathrm{FP}(\tau)} = \frac{\sum \mathrm{TP}(\tau)}{\mathrm{all\ detections}(\tau)}$ [b]
Recall(τ)	$\mathrm{Recall}(\tau) = \frac{\sum \mathrm{TP}(\tau)}{\sum \mathrm{TP}(\tau)+\sum \mathrm{FN}(\tau)} = \frac{\sum \mathrm{TP}(\tau)}{\mathrm{all\ ground\ truths}}$ [b]
AP	Area under a pre-processed Precision(τ) × Recall(τ) curve[c]
mAP	The average of AP of all object categories evaluated
AR	The average of the maximum obtained Recall across several IoU thresholds
mAR	The average of AR of all object categories evaluated
$\mathrm{AP}^{\mathrm{bbox}}$	AP at IoU = 0.50 : 0.05 : 0.95 for object detection, i.e., drawing bboxes of detected objects.
$\mathrm{AP}^{\mathrm{bbox}}_{0.50}$	AP at IoU = 0.50 (PASCAL VOC metric) for object detection.
$\mathrm{AP}^{\mathrm{bbox}}_{0.75}$	AP at IoU = 0.75 (strict metric) for object detection.
$\mathrm{AP}^{\mathrm{segm}}$	AP at IoU = 0.50 : 0.05 : 0.95 for instance segmentation, i.e., generating individual masks of detected objects.
$\mathrm{AP}^{\mathrm{segm}}_{0.50}$	AP at IoU = 0.50 (PASCAL VOC metric) for instance segmentation.
$\mathrm{AP}^{\mathrm{segm}}_{0.75}$	AP at IoU = 0.75 (strict metric) for instance segmentation.
$\mathrm{AR}^{\mathrm{bbox}}_{1}$	AR given 1 detections per image.
$\mathrm{AR}^{\mathrm{bbox}}_{10}$	AR given 10 detections per image.
$\mathrm{AR}^{\mathrm{bbox}}_{100}$	AR given 100 detections per image.
$\mathrm{AR}^{\mathrm{segm}}_{1}$	AR given 1 detections per image. segmentation.
$\mathrm{AR}^{\mathrm{segm}}_{10}$	AR given 10 detections per image.
$\mathrm{AR}^{\mathrm{segm}}_{100}$	AR given 100 detections per image.
$\mathrm{AIoU}_{0.50}$	Instance-wise average IoU at the threshold 0.50.
$\mathrm{AIoU}_{0.75}$	Instance-wise average IoU at the threshold 0.75.

[a] GT represents *ground truth*, P represents *prediction*; GT and P can represent either bboxes or binary masks.

[b] τ is the confidence threshold for filtering predictions.

[c] Microsoft COCO evaluation metrics use 101-point interpolation to calculate the AP, while the PASCAL VOC challenge uses 11-point or all-point interpolation.

5.2. Qualitative results

The qualitative results of adherent cell instance segmentation are presented in Fig. 9. At first glance, Mask R-CNN, Mask Scoring R-CNN, PointRend, and QueryInst can apparently identify individual cells relatively well, as echoed by their numerical results in Table 3. However, their segmentation details are unsatisfactory. A typical problem is that a few titled cells are always neglected, particularly when cells are densely distributed.

Additionally, it appears that ANCIS, and Rotated Mask R-CNN exhibit less satisfactory performance in comparison. This could be attributed to several reasons: ANCIS incorporates attention mechanisms in both object detection and post-detection segmentation stages, but its object detection results are somewhat inadequate, leading to less accurate segmentation. Similarly, Rotated Mask R-CNN adopts rotated bboxes for detection, but its experimental outcomes still fall short of our method's performance.

Interestingly, StarDist performs well in the detection and segmentation of healthy cells. There are essentially no fragmented masks. This is likely due to its use of convex polygons to match cell contours, naturally predicting each cell to be of an approximate elliptical shape. However, its drawbacks are also evident. Once it encounters cells that cannot be matched with convex polygons, it becomes powerless. This can also be

Table 3
Performance of several instance segmentation algorithms on our dataset.

Method	Backbone	Category	AP^{bbox}	$AP^{bbox}_{0.50}$	$AP^{bbox}_{0.75}$	AR^{bbox}_{1}	AR^{bbox}_{10}	AR^{bbox}_{100}	AP^{segm}	$AP^{segm}_{0.50}$	$AP^{segm}_{0.75}$	AR^{segm}_{1}	AR^{segm}_{10}	AR^{segm}_{100}	$AIoU_{0.50}$	$AIoU_{0.75}$
Mask R-CNN	R-50	Cells	0.536	0.910	0.578	0.038	0.305	0.610	0.503	0.880	0.536	0.036	0.288	0.570		
		Sick cells	0.187	0.417	0.120	0.095	0.301	0.334	0.190	0.419	0.153	0.094	0.293	0.323		
		W. Avg.	0.495	**0.851**	0.524	0.045	0.305	0.577	0.466	0.825	0.490	0.043	0.289	0.541	0.760	0.833
	R-101	Cells	0.552	0.904	0.617	0.039	0.313	0.622	0.508	0.867	0.576	0.036	0.294	0.573		
		Sick cells	0.194	0.411	0.154	0.097	0.312	0.36	0.189	0.431	0.152	0.096	0.298	0.336		
		W. Avg.	**0.509**	0.845	0.562	0.046	**0.313**	**0.591**	0.470	0.815	0.526	0.043	0.294	0.545	0.773	0.836
Mask Scoring R-CNN	R-50	Cells	0.517	0.899	0.554	0.038	0.296	0.591	0.501	0.878	0.543	0.037	0.294	0.550		
		Sick cells	0.179	0.411	0.134	0.101	0.292	0.333	0.187	0.443	0.128	0.096	0.298	0.327		
		W. Avg.	0.477	0.841	0.504	0.045	0.296	0.560	0.464	0.826	0.494	0.044	0.294	0.523	0.746	0.831
	R-101	Cells	0.532	0.903	0.588	0.038	0.306	0.606	0.503	0.862	0.557	0.037	0.294	0.550		
		Sick cells	0.172	0.376	0.131	0.094	0.292	0.328	0.184	0.411	0.155	0.099	0.298	0.321		
		W. Avg.	0.489	0.840	0.534	0.045	0.304	0.573	0.465	0.808	0.509	0.044	0.294	0.523	0.762	0.831
PointRend	R-50	Cells	0.541	0.905	0.595	0.037	0.305	0.613	0.539	0.902	0.626	0.037	0.297	0.598		
		Sick cells	0.208	0.432	0.183	0.105	0.315	0.356	0.200	0.442	0.185	0.104	0.303	0.343		
		W. Avg.	0.501	0.849	0.546	0.045	0.306	0.582	0.499	0.847	0.574	**0.045**	0.298	0.568	0.766	0.831
QueryInst	R-50	Cells	0.329	0.673	0.276	0.028	0.211	0.548	0.355	0.671	0.351	0.029	0.222	0.559		
		Sick cells	0.072	0.186	0.044	0.064	0.230	0.342	0.071	0.184	0.048	0.065	0.227	0.338		
		W. Avg.	0.298	0.615	0.248	0.032	0.213	0.524	0.321	0.613	0.315	0.033	0.223	0.533	0.761	0.832
	R-101	Cells	0.386	0.729	0.370	0.033	0.237	0.578	0.408	0.727	0.431	0.034	0.243	0.581		
		Sick cells	0.087	0.234	0.050	0.069	0.241	0.358	0.078	0.207	0.046	0.067	0.232	0.344		
		W. Avg.	0.350	0.670	0.332	0.037	0.237	0.552	0.369	0.665	0.385	0.038	0.242	0.553	0.771	0.834
Rotated Mask R-CNN	R-50	Cells	0.337	0.841	0.168	0.026	0.212	0.483	0.417	0.816	0.388	0.033	0.255	0.513		
		Sick cells	0.139	0.373	0.069	0.089	0.244	0.301	0.172	0.416	0.096	0.098	0.257	0.314		
		W. Avg.	0.313	0.785	0.156	0.033	0.216	0.461	0.388	0.768	0.353	0.041	0.255	0.489	0.698	0.815
	R-101	Cells	0.345	0.837	0.187	0.026	0.216	0.490	0.429	0.828	0.404	0.032	0.257	0.539		
		Sick cells	0.026	0.216	0.490	0.077	0.241	0.295	0.157	0.406	0.082	0.085	0.240	0.297		
		W. Avg.	0.307	0.763	0.223	0.032	0.219	0.467	0.397	0.778	0.366	0.038	0.255	0.510	0.701	0.820
ANCIS	R-50	Cells	0.360	0.669	0.346	0.032	0.265	0.426	0.206	0.488	0.136	0.017	0.187	0.316		
		Sick cells	0.140	0.348	0.070	0.089	0.233	0.235	0.114	0.292	0.060	0.074	0.201	0.204		
		W. Avg.	0.334	0.631	0.313	0.039	0.261	0.403	0.195	0.465	0.127	0.024	0.189	0.303	0.740	0.823
	R-101	Cells	0.387	0.690	0.403	0.033	0.282	0.454	0.218	0.501	0.150	0.018	0.188	0.330		
		Sick cells	0.167	0.351	0.131	0.089	0.290	0.296	0.133	0.330	0.079	0.074	0.243	0.249		
		W. Avg.	0.361	0.650	0.371	0.040	0.283	0.435	0.208	0.481	0.142	0.025	0.195	0.320	0.744	0.823
Oriented R-CNN	R-50	Cells	0.215	0.775	0.039	0.046	0.272	0.220	0.549	0.917	0.636	0.031	0.320	0.593		
		Sick cells	0.070	0.273	0.009	0.031	0.067	0.087	0.180	0.414	0.131	0.003	0.082	0.244		
		W. Avg.	0.198	0.715	0.035	0.044	0.248	0.204	0.505	0.857	0.576	0.028	0.292	0.552	0.717	0.826
	R-101	Cells	0.245	0.813	0.066	0.113	0.295	0.252	0.562	0.930	0.667	0.008	0.342	0.605		
		Sick cells	0.083	0.311	0.017	0.019	0.080	0.100	0.196	0.438	0.168	0.005	0.089	0.269		
		W. Avg.	0.226	0.753	0.060	**0.102**	0.269	0.234	0.518	**0.872**	0.608	0.008	0.312	0.565	0.728	0.828
StarDist	U-Net	Cells	0.150	0.493	0.047	0.016	0.129	0.280	0.334	0.796	0.170	0.026	0.208	0.421		
		Sick cells	0.059	0.172	0.031	0.045	0.113	0.113	0.078	0.207	0.044	0.052	0.124	0.125		
		W. Avg.	0.139	0.455	0.045	0.019	0.127	0.260	0.304	0.726	0.155	0.029	0.198	0.386	0.722	0.805
Our method	R-50	Cells	0.546	0.885	0.611	0.037	0.310	0.618	0.591	0.897	0.733	0.037	0.318	0.643		
		Sick cells	0.223	0.441	0.237	0.103	0.290	0.294	0.219	0.431	0.228	0.101	0.280	0.284		
		W. Avg.	**0.508**	0.832	**0.567**	0.045	**0.308**	0.579	**0.547**	0.842	**0.673**	**0.045**	**0.313**	**0.600**	**0.819**	**0.848**
	R-101	Cells	0.535	0.883	0.599	0.037	0.307	0.609	0.582	0.896	0.720	0.037	0.315	0.634		
		Sick cells	0.228	0.446	0.207	0.109	0.306	0.311	0.225	0.443	0.220	0.103	0.297	0.303		
		W. Avg.	0.498	0.831	0.552	0.046	0.307	0.574	0.540	0.842	0.661	**0.045**	**0.313**	0.595	0.816	0.846

The overall best AP scores for both healthy and unhealthy cells are highlighted in bold. Specifically, our method achieves a weighted average of 0.673 in $AP^{segm}_{0.75}$, indicating strong performance under the most stringent criteria for TP identification. While our method's performance may be slightly less competitive in some other metrics, it stands out in capturing nearly every cell, as shown in Fig. 9.

observed from Table 3, where its performance is not as satisfactory for unhealthy cells, which exhibit diverse shapes.

Overall, our method demonstrates improvement. It can accurately detect and segment cells in sparsely and densely distributed situations.

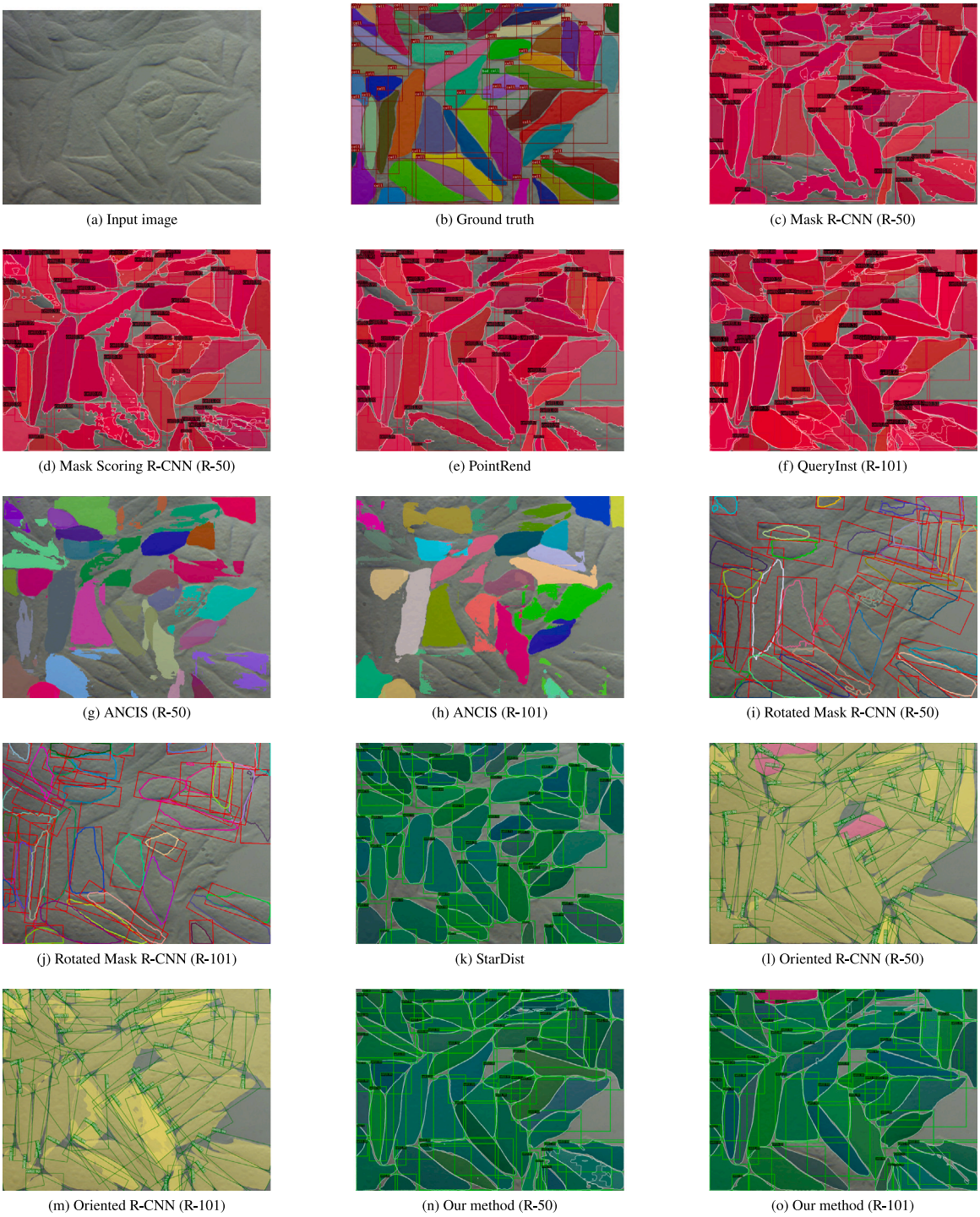
6. Discussion

6.1. Value of the proposed dataset for image cytometry and microinjection

Our dataset curated in this study can be valuable for image cytometry, which involves the quantitative analysis of cellular properties and processes at the single-cell level, such as cell cycle analysis, cell morphology characterization, or cellular feature extraction. Our dataset can also be used in cell microinjection, where specific substances or materials are injected into individual cells for various experimental purposes.

6.2. Limitations of the proposed dataset

Our dataset has several limitations that need to be considered. Firstly, the dataset is relatively small, which may restrict the diversity and variability of the available data for training and evaluation. Secondly, the dataset only includes images of a single cell line, namely HepG2 human liver cancer cells. This limitation raises concerns about the generalizability of the proposed method to other cell types, as different cells may exhibit distinct characteristics and segmentation challenges. Lastly, the dataset comprises images captured at a single magnification, lacking different imaging conditions and resolutions



(a) Input image (b) Ground truth (c) Mask R-CNN (R-50)

(d) Mask Scoring R-CNN (R-50) (e) PointRend (f) QueryInst (R-101)

(g) ANCIS (R-50) (h) ANCIS (R-101) (i) Rotated Mask R-CNN (R-50)

(j) Rotated Mask R-CNN (R-101) (k) StarDist (l) Oriented R-CNN (R-50)

(m) Oriented R-CNN (R-101) (n) Our method (R-50) (o) Our method (R-101)

Fig. 9. Qualitative results of instance segmentation for a sample adherent cell image. Our method effectively captures nearly every cell, demonstrating high accuracy and robustness in identifying and delineating individual cells.

commonly encountered in other biomedical scenarios. We recognize these limitations and plan to address them in future work by expanding the dataset to include additional cell lines and imaging conditions to enhance generalizability.

6.3. Limitations of the proposed method

While our method is promising, it does have certain limitations that should be improved. Firstly, it is not a fully end-to-end trainable network, necessitating multi-step training and inference procedures. Secondly, its deployment as a real-time solution may pose challenges due to the relatively extended inference process. Thirdly, the method itself retains a dependence on bbox design and RPN, aspects that might be considered somewhat intricate.

We also acknowledge the current weak performance for classification, as there are minimal stable characteristics to distinguish between them. It is worth noting that only increasing the dataset size is crucial for introducing more information into the classification process. This

is also our ongoing work; we are also trying to distinguish them by introducing other information on this matter.

We know our method is slower and requires more disk space even without quantitative measurements. However, the primary goal of developing this method was to achieve high accuracy and superior visualization results, thus we made some compromises in terms of efficiency. Specifically, our method consists of one pre-processing step and three major steps: pre-processing by rotating images and annotations to create an interim dataset, training a Mask R-CNN, training a ResNet classifier, and finally, predicting and evaluating. While the dataset processing and network training are time-efficient and comparable to other CNNs, the most time-consuming part is creating the interim dataset for ResNet training. At this stage, every prediction produced by Mask R-CNN must be compared against the ground truth to determine if it should be kept or discarded, which involves three cascaded loops and takes several hours due to non-optimized code. Additionally, the mask selection stage in Step 3 also contains three cascaded loops, taking approximately one hour. As it stands, our method is not yet suitable for deployment due to these inefficiencies, but we plan to optimize the algorithm and reduce computational complexity in future work.

6.4. Comparison with oriented R-CNN and TTA

When comparing our method with Oriented R-CNN, it can be observed that at an IoU threshold of 0.50, Oriented R-CNN's performance is not bad, only slightly worse than our method. However, at an IoU threshold of 0.75, its performance lags behind our method. Oriented R-CNN is an innovation based on Mask R-CNN, where the innovation lies in using oriented bbox proposals instead of the standard bbox proposals of Mask R-CNN. It generates arbitrarily oriented bbox proposals by increasing the RPN regression branch output parameters from four to six, using a midpoint offset representation scheme. These oriented bboxes are more effective for various inclined objects in remote sensing images, such as tilted parked airplanes and ships. However, aside from the bbox proposal, Oriented R-CNN and Mask R-CNN share the same subsequent steps for mask training, prediction, and filtering.

In contrast, our method is independently designed for cell instance segmentation. We simply rotate the images and ground truth by 45° and feed them into Mask R-CNN, where the resulting bbox proposals are still traditional rectangular bboxes with only four output parameters. This approach does not remarkably increase the memory requirements of a graphics processing unit (GPU) during network training.

Our method, on the other hand, not only uses DVI for training (which serves a similar function to the oriented bbox proposal) but also abandons the NMS of bbox used by Mask R-CNN. Instead of first determining whether a bbox is a valid prediction and then comparing the mask within this bbox with the ground truth, we directly determine whether a mask is a valid prediction and compare this mask directly with the ground truth. This approach makes our method superior in terms of mask details, particularly at an IoU threshold of 0.75, where a mask prediction is only considered valid if its IoU with the ground truth exceeds 0.75. Our method performs the best at this threshold.

While the DVI part in our method and TTA share similarities, our method is not only DVI, the SMS also plays a vital role in improving our method's overall performance. The difference can be summarized as follows: TTA focuses on applying data augmentation during the test stage to enhance model performance, whereas our method employs DVI during both training and testing to leverage additional information from different perspectives. Additionally, our approach introduces an SMS component, which is absent in traditional TTA. This SMS component directly processes the mask predictions from dual-view testing, aiming to improve the final outputs by selectively refining the results.

6.5. Applicability to other domains and future work

We believe that this DVI and SMS can be applied to many biomedical imaging tasks for object detection and instance segmentation, as long as there are objects with various angles of inclination in those biomedical images, such as cytology or histopathology images.

Cytology involves the study of individual cells, analyzing their structure, function, and behavior, and plays a critical role in cancer screening and diagnosis. For instance, the Pap smear is a common cytological procedure used to examine cervical cells for signs of cervical cancer or abnormalities. Due to cell overlap in cervical samples, researchers have devoted special attention to addressing this challenge [39,40].

Histopathology is a subset of pathology focused on microscopic examination of tissues to diagnose diseases. In this context, researchers search for cellular changes indicative of specific diseases in microscope images of tissue samples obtained from surgeries or biopsies. For instance, characteristic Reed–Sternberg cells in a lymph node tissue sample may lead to a diagnosis of Hodgkin's lymphoma. Tissues are typically stained and fixed, meaning that the cells are dead, and fluorescence microscopy is often used for imaging [41]. One recent example is Mesmer [14]. However, it is important to note that histopathology images differ from PhC and DIC images, even though all can be referred to as cell images. The researchers' focus and the emphasis on the developed algorithms vary accordingly. Moreover, the metrics for semantic segmentation and instance segmentation also differ. We hope to include these specific algorithms for comparison in our future work.

7. Conclusion

In this study, we created a dataset comprising 520 DIC images of 12,198 unstained living HepG2 human liver cancer cells. Each DIC image was paired with a corresponding fluorescence image stained with calcein AM, ensuring high-quality annotations for ground-truth labeling. Our dataset is a pioneer in considering the multi-state nature of adherent cells, where both healthy and unhealthy cells can be observed within a single image.

We developed a new cascaded method for instance segmentation of unstained living adherent cells. Unlike existing approaches that primarily support single-state cell detection and segmentation, our method natively handles multi-state (multi-category) cell detection and instance segmentation. The experimental results demonstrate that our method achieves a weighted average of 0.567 in $AP^{bbox}_{0.75}$ and a weighted average of 0.673 in $AP^{segm}_{0.75}$. Such an advantage can be attributed to two novel methods, i.e., DVI to combine original and rotated views as input to capture cell instances as much as possible and SMS to select the finest instances in a supervised manner.

Overall, our study contributes a unique multi-state cell image dataset and a novel method for cell detection and instance segmentation. Through comparisons, we demonstrate the effectiveness of using rotated images and supervised mask refinement in our method. Our work could be of broad interest to researchers and stimulate new ideas in biological image analysis.

Code availability

The code generated during the current study is available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Fei Pan: Writing – review & editing, Writing – original draft, Project administration, Formal analysis, Data curation, Conceptualization. **Yutong Wu:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis. **Kangning Cui:** Writing – original draft, Methodology, Investigation.

Shuxun Chen: Resources. **Yanfang Li:** Resources. **Yaofang Liu:** Writing – original draft, Methodology. **Adnan Shakoor:** Resources. **Han Zhao:** Resources, Data curation. **Beijia Lu:** Validation. **Shaohua Zhi:** Writing – original draft. **Raymond Hon-Fu Chan:** Supervision, Funding acquisition. **Dong Sun:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset will be accessible via the following link: https://doi.org/10.5281/zenodo.13120679.

Declaration of Generative AI and AI-assisted technologies in the writing process

Statement: During the preparation of this work the authors used ChatGPT to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgments

This work was supported in part by the InnoHK Project on Project 2.6 — "Magneto/optical steered vascular microrobotic system for image-guided CVD intervention" at the Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE). It was also supported by Grants 11211421, CityU11301120, CityU11309922, and C1013-21GF from the Research Grants Council of Hong Kong. Additional support was provided by Grant 9380101 from the City University of Hong Kong. The viewpoints expressed herein are solely those of the authors and do not represent the opinions of InnoHK — ITC, the Research Grants Council of Hong Kong, or the City University of Hong Kong. This work was partially conducted by using the computational facilities, CityU Burgundy, managed and provided by the Computing Services Center at the City University of Hong Kong.

References

[1] B. Alberts, D. Bray, K. Hopkin, A.D. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Essential Cell Biology, 4th, Garland Science, New York, NY, 2013.
[2] E. Moen, D. Bannon, T. Kudo, W. Graf, M. Covert, D.V. Valen, Deep learning for cellular image analysis, Nature Methods 16 (12) (2019) 1233–1246, doi:10/gf26n8.
[3] T. Vicar, J. Balvan, J. Jaros, F. Jug, R. Kolar, M. Masarik, J. Gumulec, Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison, BMC Bioinformatics 20 (2019) 360, doi:10/ggtqhb.
[4] G. Karp, J. Iwasa, W. Marshall, Karp's Cell and Molecular Biology: Concepts and Experiments, eighth ed., Wiley, 2015.
[5] J. Icha, M. Weber, J.C. Waters, C. Norden, Phototoxicity in live fluorescence microscopy, and how to avoid it, BioEssays 39 (8) (2017) 1700003, doi:10/gbphfn.
[6] E. Meijering, Cell segmentation: 50 years down the road, IEEE Signal Process. Mag. 29 (5) (2012) 140–145, doi:10/gfghgs.
[7] F. Pan, Y. Jiao, S. Chen, L. Xing, D. Sun, Deep learning-enhanced dual-module large-throughput microinjection system for adherent cells, IEEE Trans. Autom. Sci. Eng. 20 (4) (2023) 2409–2422, doi:gq5fhz.
[8] J. Liu, V. Siragam, Z. Gong, J. Chen, M.D. Fridman, C. Leung, Z. Lu, C. Ru, S. Xie, J. Luo, R.M. Hamilton, Y. Sun, Robotic adherent cell injection for characterizing cell-cell communication, IEEE Trans. Biomed. Eng. 62 (1) (2014) 119–125, doi:10/f6tpxw.
[9] M. Maška, V. Ulman, P. Delgado-Rodriguez, E. Gómez-de-Mariscal, T. Nečasová, F.A. Guerrero Peña, T.I. Ren, E.M. Meyerowitz, T. Scherr, K. Löffler, R. Mikut, T. Guo, Y. Wang, J.P. Allebach, R. Bao, N.M. Al-Shakarji, G. Rahmon, I.E. Toubal, K. Palaniappan, F. Lux, P. Matula, K. Sugawara, K.E.G. Magnusson, L. Aho, A.R. Cohen, A. Arbelle, T. Ben-Haim, T.R. Raviv, F. Isensee, P.F. Jäger, K.H. Maier-Hein, Y. Zhu, C. Ederra, A. Urbiola, E. Meijering, A. Cunha, A. Muñoz-Barrutia, M. Kozubek, C. Ortiz-de-Solórzano, The cell tracking challenge: 10 years of objective benchmarking, Nature Methods 20 (7) (2023) 1010–1020, doi:10/gskzv4.
[10] S. Baar, M. Kuragano, K. Tokuraku, S. Watanabe, Towards a comprehensive approach for characterizing cell activity in bright-field microscopic images, Sci. Rep. 12 (1) (2022) 16884, doi:10/grnfbg.
[11] W. Wang, D. Douglas, J. Zhang, S. Kumari, M.S. Enuameh, Y. Dai, C.T. Wallace, S.C. Watkins, W. Shu, J. Xing, Live-cell imaging and analysis reveal cell phenotypic transition dynamics inherently missing in snapshot data, Sci. Adv. 6 (36) (2020) eaba9319, doi:10/gj6tnw.
[12] L. Maddalena, L. Antonelli, A. Albu, A. Hada, M.R. Guarracino, Artificial intelligence for cell segmentation, event detection, and tracking for label-free microscopy imaging, Algorithms 15 (9) (2022) 313, doi:10/grnd94.
[13] C. Edlund, T.R. Jackson, N. Khalid, N. Bevan, T. Dale, A. Dengel, S. Ahmed, J. Trygg, R. Sjögren, LIVEcell—A large-scale dataset for label-free live cell segmentation, Nature Methods 18 (9) (2021) 1038–1045, doi:10/gmptqs.
[14] N.F. Greenwald, G. Miller, E. Moen, A. Kong, A. Kagel, T. Dougherty, C.C. Fullaway, B.J. McIntosh, K.X. Leow, M.S. Schwartz, C. Pavelchek, S. Cui, I. Camplisson, O. Bar-Tal, J. Singh, M. Fong, G. Chaudhry, Z. Abraham, J. Moseley, S. Warshawsky, E. Soon, S. Greenbaum, T. Risom, T. Hollmann, S.C. Bendall, L. Keren, W. Graf, M. Angelo, D. Van Valen, Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning, Nature Biotechnol. 40 (4) (2022) 555–565, doi:10/gnm9wv.
[15] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2961–2969, doi:10/gfghjd.
[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi:10/gdcfkn.
[17] S. Minaee, Y.Y. Boykov, F. Porikli, A.J. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 44 (7) (2021) 3523–3542, doi:10/gjh26c.
[18] A.M. Hafiz, G.M. Bhat, A survey on instance segmentation: state of the art, Int. J. Multimed. Inf. Retr. 9 (3) (2020) 171–189, doi:10/gg5g23.
[19] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[20] Z. Huang, L. Huang, Y. Gong, C. Huang, X. Wang, Mask scoring R-CNN, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6409–6418, doi:10/gg3d6j.
[21] S. Looi, Rotated mask R-CNN: from bounding boxes to rotated bounding boxes, 2019, doi:10/kpdq.
[22] X. Xie, G. Cheng, J. Wang, X. Yao, J. Han, Oriented R-CNN for object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3520–3529, doi:10/grkrgv.
[23] X. Xie, G. Cheng, J. Wang, K. Li, X. Yao, J. Han, Oriented R-CNN and beyond, Int. J. Comput. Vis. (2024) doi:10/gtn434.
[24] A. Kirillov, Y. Wu, K. He, R. Girshick, PointRend: image segmentation as rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9799–9808.
[25] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, W. Liu, Instances as queries, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 6910–6919, doi:10/gqggpx.
[26] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Proceedings of the European Conference on Computer Vision (ECCV), in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 213–229, doi:10/ghw3xx.
[27] U. Schmidt, M. Weigert, C. Broaddus, G. Myers, Cell detection with star-convex polygons, in: A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 265–273, doi:10/ggnzqb.
[28] J.C. Caicedo, A. Goodman, K.W. Karhohs, B.A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, A.E. Carpenter, Nucleus segmentation across imaging experiments: the 2018 data science bowl, Nature Methods 16 (12) (2019) 1247–1253, doi:10/ggcd7h.
[29] J. Yi, P. Wu, M. Jiang, Q. Huang, D.J. Hoeppner, D.N. Metaxas, Attentive neural cell instance segmentation, Med. Image Anal. 55 (2019) 228–240, doi:10/gg73zt.
[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.), Proceedings of the European Conference on Computer Vision (ECCV), in: Lecture Notes in Computer Science, Springer International Publishing, Amsterdam, The Netherlands, 2016, pp. 21–37, doi:10/gc7rk8.
[31] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, in: Lecture Notes in Computer Science, 9351, Springer International Publishing, Munich, Germany, 2015, pp. 234–241, doi:10/gcgk7j.
[32] W. Wang, D.A. Taft, Y.-J. Chen, J. Zhang, C.T. Wallace, M. Xu, S.C. Watkins, J. Xing, Learn to segment single cells with deep distance estimator and deep cell detector, Comput. Biol. Med. 108 (2019) 133–141, doi:10/gg73zv.

[33] T. Prangemeier, C. Reich, H. Koeppl, Attention-based transformers for instance segmentation of cells in microstructures, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020, pp. 700–707, doi:10/gkcqx7.
[34] N. Moshkov, B. Mathe, A. Kertesz-Farkas, R. Hollandi, P. Horvath, Test-time augmentation for deep learning-based cell segmentation on microscopy images, Sci. Rep. 10 (1) (2020) 5068, doi:10/ggsnnk.
[35] K. Nishimura, C. Wang, K. Watanabe, D.F.E. Ker, R. Bise, Weakly supervised cell instance segmentation under various conditions, Med. Image Anal. 73 (2021) 102182, doi:10/gm8gkw.
[36] C. Stringer, T. Wang, M. Michaelos, M. Pachitariu, Cellpose: A generalist algorithm for cellular segmentation, Nature Methods 18 (1) (2021) 100–106, doi:10/ghrgms.
[37] J. Yi, P. Wu, H. Tang, B. Liu, Q. Huang, H. Qu, L. Han, W. Fan, D.J. Hoeppner, D.N. Metaxas, Object-guided instance segmentation with auxiliary feature refinement for biological images, IEEE Trans. Med. Imaging 40 (9) (2021) 2403–2414, doi:10/gm8ggf.
[38] K. Wada, Labelme: image polygonal annotation with python, 2016, URL: https://github.com/wkentaro/labelme.
[39] J. Zhao, Y.-J. He, S.-Q. Zhao, J.-J. Huang, W.-M. Zuo, AL-net: attention learning network based on multi-task learning for cervical nucleus segmentation, IEEE J. Biomed. Health Inf. 26 (6) (2022) 2693–2702, doi:10/gshdp9.
[40] Y. Zhou, H. Chen, H. Lin, P.-A. Heng, Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation, in: A.L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M.A. Zuluaga, S.K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 521–531, doi:10/gmqkrc.
[41] E.S. Nasir, A. Parvaiz, M.M. Fraz, Nuclei and glands instance segmentation in histology images: A narrative review, Artif. Intell. Rev. 56 (8) (2023) 7909–7964, doi:10/gshdpx.