# Blurry-Consistency Segmentation Framework with Selective Stacking on Differential Interference Contrast 3D Breast Cancer Spheroid

Thanh-Huy Nguyen, Thi Kim Ngan Ngo, Mai Anh Vu, Ting-Yuan Tu

IMBSL Lab, Department of Biomedical Engineering

National Cheng Kung University, Tainan, Taiwan ROC

harvey.nguyen0811@gmail.com

## Abstract

*The ability of three-dimensional (3D) spheroid modeling to study the invasive behavior of breast cancer cells has drawn increased attention. The deep learning-based image processing framework is very effective at speeding up the cell morphological analysis process. Out-of-focus photos taken while capturing 3D cells under several z-slices, however, could negatively impact the deep learning model. In this work, we created a new algorithm to handle blurry images while preserving the stacked image quality. Furthermore, we proposed a unique training architecture that leverages consistency training to help reduce the bias of the model when dense-slice stacking is applied. Additionally, the model's stability is increased under the sparse-slice stacking effect by utilizing the self-training approach. The new blurring stacking technique and training flow are combined with the suggested architecture and self-training mechanism to provide an innovative yet easy-to-use framework. Our methods produced noteworthy experimental outcomes in terms of both quantitative and qualitative aspects.*

## 1. Introduction

Recent advancements in the utilization of 3D multicellular tumor spheroids have significantly progressed research in oncogenesis and tissue engineering, enriching our understanding of these domains [29]. While various software tools exist for analyzing spheroid behaviors using microscopic data, they are limited to extracting specific cell morphologies at specific time points during time-lapse observation [6, 7, 9, 15]. Differential interference contrast (DIC) microscopy, with its fluorescence-free and non-destructive detection advantages, emerges as a widely preferred approach for observing live cells and tracking their dynamic progression temporally [34].

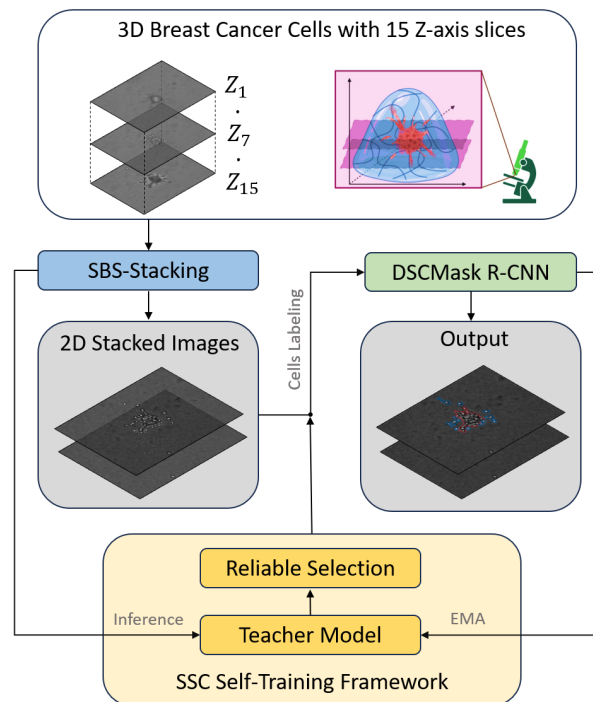Despite its advantages, analyzing DIC spheroid images



Figure 1. Our proposed pipeline for stacking and training 3D cell images. The 3D microscopy images are, firstly, stacked before running through the training stage of the proposed architecture DSCMask R-CNN. Finally, a self-training framework is employed to select reliable samples for the continuous training epochs.

remains time-consuming and poses challenges in automating model extraction and analysis[15]. The introduction of supervised machine learning frameworks, particularly deep convolutional neural networks (CNNs), has revolutionized medical image analysis [16, 17, 27]. Various deep learning-based variants, such as encoder-decoder structures like U-net or two-stage instance segmentation like Mask R-CNN, leverage contextual information for image segmentation. Mask R-CNN model is an instance segmentation architecture leveraging the power of three heads, making them

promising for analyzing spheroid behavior in DIC bright-field images.

Prior methods for 3D cell multiple slices have been extensively studied [22]. Most of these methods use a variety of ways to transform the source image into different scales in an effort to improve resolution. However, traditional image quality enhancement from digital image processing approaches no longer shows good performance compared to deep learning approaches. The previous deep learning method [25] learns focus measures to identify focused and defocused pixels in the source using an encoder-decoder-based approach. With the growth of consistency-based learning methods [], segmentation models have been assisted to perform excellently under noisy input, such as hard-to-learn samples.

Having said that, most of those approaches are not well-utilized on 3D DIC microscope images due to the heavy blur effect of multiple slices. Focus Stacking itself showed promising performance with fast inference time and less sensitivity against the blur effect. In this study, our objective is to adapt the existing Focus Stacking framework, introducing a modified version termed Selective Blurry-Slice Stacking (SBS-Stacking) by filtering a high percentage of blurred areas to stack images before stacking. Besides, inspired by FixMatch [23], we redesigned Mask R-CNN by adding a weak-to-strong augmentation mechanism to leverage the well-segmented sample from SBS-Stacking to teach fully-stacked blur images, called dense-stacked samples. Finally, to select useful information from stacked images with fewer slices, called sparse-stacked samples, we proposed a new self-training module. Different from dense-stacked information assisted by the consistency module, sparse-stacked information is utilized by a sparse-stacking consistency self-training framework that can select and prioritize reliable samples for re-training.

Our main contributions are summarized as follows:

- We proposed SBS-stacking methods to select high-quality slices for labeling and training procedures.
- Inspired by semi-supervised-based pseudo-labeling methods, we redesigned Mask R-CNN that combined consistency regularization loss, called DSCMask R-CNN to leverage dense-stacked information from all slices.
- We proposed a novel self-training scheme that utilizes sparse-stacked information that performs selective re-training via prioritizing reliable images.

## 2. Related Works

### 2.1. Biomedical Image Stacking

Previous approaches for 3D cell multiple slices have been researched widely [22]. These approaches fall into three categories: spatial domain (pixel values), transform domain (frequency components), and deep learning methods. Spa-

tial domain image fusion selects pixels or regions from focused parts [3]. Transform domain methods enhance resolution by transforming the source image into different scales, employing techniques like wavelet, curvelet, contourlet transforms, neighbor distance, Laplacian pyramid, or gradient pyramid [10]. Deep learning methods learn focus measures to identify focused and defocused pixels in source images [25], fusion operations to combine images without ground truth [20], and direct mappings between high and low-frequency images [12].

### 2.2. Consistency Regularization

Pseudo-labeling [2] and consistency regularization [4, 14] are two powerful strategies for exploiting unlabelled data. In the conventional pseudo-labeling method, the predicted pixels with high confidence are chosen as the pseudo-label for the unlabeled data using a high and defined threshold [11]. However, only a small number of samples can surpass the selected threshold at the beginning of training with this technique. Consequently, Zhang et al. [33] introduce the Curriculum Pseudo Labeling (CPL) technique, which modifies the flexible threshold of every category in real-time while it is being trained. Low-confidence pixels will nevertheless be eliminated in spite of the favorable results. Low-condence pixels are crucial for model training, as Wang et al. [28] showed. Consistency regularization seems to use data more, therefore this study will concentrate more on it. Consistency regularization concentrates more on how to obtain two identical predictions, such as data perturbation [8], model perturbation [13], feature perturbation [18], etc., even if it makes use of all available prediction data. However, when it comes to prediction optimization, a consistency loss function, such as L2 Loss, is used to treat all data consistently.

### 2.3. Self-training on Semantic Segmentation

The most basic pseudo-labeling and semi-supervised techniques are called self-training methods; they were initially introduced in [31], extensively examined in [26], and used with deep neural networks for the first time in [11]. These techniques involve feeding back the training set with the model's predictions, so retraining a base supervised model. It has been receiving more and more attention lately from a variety of domains, including domain adaptation [35], semi-supervised learning [5], and fully-supervised image recognition [19]. Specifically, it has been reconsidered in several semi-supervised tasks, such as semantic segmentation [32], object detection [24], and picture classification [5]. J.Yuan et al. [32] base their method on the premise that excessive data augmentations are detrimental to clean data distribution. Several methods for integrating data augmentation techniques into the self-training process have also been suggested. During the self-training phase, the ST++ [30]

method uses data augmentation techniques on the unlabeled images.

## 3. Methodology

### 3.1. Selective Blurry-Slice Stacking (SBS-Stacking)

First, we derive a blur map for an image $\mathbf{B}$, characterizing spatially varying blur by modeling it as a composite of a sharp image ($\mathbf{S}$) convolved with a blur kernel ($\mathbf{b}$) and the addition of noise $\mathbf{n}$, as demonstrated in[1].

$$\mathbf{B} = \mathbf{S} \odot \mathbf{b} + \mathbf{n} \qquad (1)$$

The blur detection methodology involves the following key steps:

**Gaussian Filtering for Noise Reduction**: We employ a Gaussian filter, characterized by a small kernel function $(g(x, y))$, applied individually to each pixel (x, y) of the blurry image (B). The Gaussian filter is an objectives function defined by:

$$g(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (2)$$

This process effectively mitigates high-frequency noise (introduced by the device lens). The resulting filtered image is denoted as Bg, retaining the image's inherent shape and structure while eliminating spatial redundancy. In the experiment setting, we chose a small $\sigma = 0.5$.

**Gradient Magnitude Calculation**: Let $B_g$ represent the input blurry image after it has been Gaussian filter. Convolution operations with horizontal $(h_x)$ and vertical $(h_y)$ gradient operators are applied to the Gaussian-filtered image $(B_g)$ to compute the gradient magnitudes. The magnitude (G) is the square root of the sum of squared gradients. The and $h_x$ and $h_y$ are defined as:

$$h_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad h_y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Then, the gradient magnitude image G is calculated as:

$$G = \sqrt{(B_g \odot h_x)^2 + (B_g \odot h_y)^2}, \qquad (3)$$

These steps collectively yield a comprehensive blur map reflecting the spatially varying blur in the image, providing valuable insights into the localized variations in sharpness.

Next, we apply HiFST[1] decomposition on $G$ - the gradient-computed sample; blur can be interpreted differently depending on the scale. Images are processed with patch size M x M for each pixel centered at $(i, j)$. Given that $H_{i,j}^M$ and $L_{i,j}$ are, respectively, a vector consisting of the absolute values of the high-frequency Discrete Cosine Transform (DCT) coefficients and its sorted vector. For

each element $t$, $L_{i,j;t}$ is the $t$-th element of in vector $L_{i,j}$. The matrix $L_t$ is defined as the $N_1 \times N_2$ matrix defined as:

$$L_t = \{L_{i,j;t}, 0 \le i < N_1, 0 \le j < N_2\} \qquad (4)$$

---

**Algorithm 1** Selective Blurry-Slice Stacking (SBS-Stacking)

---

**Input:** $B$: Blur images; $k$: top-k images that in-focus; $Z$: total number of slices.
**Output:** $S$ : Sharp image.
**for** $z = 1$ to $Z$ **do**
  Calculate gradient magnitude image $G_z$ using Eq.3
  $I_z = HiFST(G_z)$ {Compute blur detection map.}
  $I_{BM}^z = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} p_{i,j}$ , $p_{i,j} \in [0, 255]$ {Compute in-focus pixel values of each $z$ image.}
  $I_Z \leftarrow I_{BM}^z$
**end for**
Sort $I_Z$ by descending order.
**for** $i = 0$ to $k - 1$ **do**
  $I_{BM}^k \leftarrow I_{BM}^z[i]$ {Select top-k slices.}
**end for**
$S = FocusStack\left(I_{BM}^k\right)$
**Return:** $S$.

---

HiFST comprised of $\frac{M^2+M}{2}$ normalized layers $\widehat{L_t}$, $1 \le t \le \frac{M^2+M}{2}$, where $\widehat{L_t}$ is shown as:

$$\widehat{L_t} = \frac{L_t - \min(L_t)}{\max(L_t) - \min(L_t)}, \quad 1 \le t \le \frac{M^2 + M}{2} \qquad (5)$$

Each layer can more effectively distinguish between the image's blurry and unblurred areas and measure the amount of blur locally by normalizing each layer between $[0, 1]$.

A blur detection map $D$ uses the first $\sum_{r=1}^{m} M_r$ layers is computed as:

$$D = T \circ \omega \qquad (6)$$

whereas $\circ$ is Hadamard product and $M_r = 2^{2+r}$ if even and $M_r = 2^{2+r} - 1$ if odd. Given that Matrices $T$ with $0 \le i < N_1, 0 \le j < N_2$, $T$ and $\omega$ elements are defined as:

$$T_{i,j} = \max\left(\left\{\widehat{L_{i,j;t}} : t = 1, \ldots, \sum_{r=1}^{m} M_r\right\}\right), \qquad (7)$$

$$\omega_{i,j} = -\sum_{(i',j') \in R_{(i,j)}^k} P(T_{i',j'}) \log[P(T_{i',j'})], \qquad (8)$$

whereas $R_{(i,j)}^k$ is $k \times k$ patch centered at pixel $(i, j)$ and $P$ is probability function. After choosing $k = 7$ for computing $7 \times 7$ neighborhood for the corresponding pixel in $T$, the

entropy map $\omega$ is used as a weighting factor to give more weight to the microscopy image's salient regions. To reduce the impact of outliers and preserve boundaries, the final blur map is smoothed with edge-preserving filters.

After achieving a full saliency map of cell images, it can be seen that some of the z-slices have a high percentage of white area, whereas the blurry slices only contain most of the black area. Many previous similar approaches[21] consider stacking all the slices, which include many noise-contained slices. The output of the focus stacking algorithm is not well-represented for identifying among single cells. We propose the SBS-Stacking algorithm that aims to select high in-focus ratio slices to stack based on the achieved saliency map of each picture.

Given that, for every $z$-slice in $Z$, cumulative in-focus pixel values of an input image are computed by using estimated blur map $I_{BM}$ with the shape of $N_1 \times N_2$ as follows:

$$I_{BM}^z = \sum_{i=0}^{N_1} \sum_{j=0}^{N_2} p_{i,j} \ , p_{i,j} \in [0, 255] \qquad (9)$$

where $p_{i,j}$ is intensity value at pixel $(i, j)$. The k-selected slice with high in-focus pixel intensity is defined as:

$$I_{BM}^k = \{top_k \left( I_{BM}^z \mid z \in Z \right), 0 < k \leq Z\} \qquad (10)$$

After selecting the k in-focus slice, the image output is focus-stacked as below:

$$S = FocusStack \left( I_{BM}^k \right) \qquad (11)$$

where FocusStack is an algorithm with the flow of using SIFT for feature detection, slices alignment with kNN feature matching, generating Gaussian blur, and Laplacian computing for combining multiple slices in the Z-axis.

The pseudo-code of the proposed stacking framework is shown in Algorithm 1.

## 3.2. Dense-Stacking Consistency Mask R-CNN (DSCMask R-CNN)

In our work, we utilize the architecture of Mask-RCNN with a strong-weak consistency mechanism under multiple-slice stacking methods. For the input, the labeled set is the output of SBS-Stacking with $k$ slices, which plays a role as weak-augmented flow. Conversely, an unlabeled dataset, generated from all $z$ slices stacked, can be seen as a strong-augmented set.

For supervised training with a labeled set, we keep the original loss of Mask-RCNN, which includes classification loss, regression loss, and supervised mask loss. The Mask-RCNN loss is computed as:

$$L_{MaskRCNN} = L_{cls} + L_{box} + L_{mask} \qquad (12)$$

where classification loss is log loss, bounding box loss is regression smooth $L1$ loss, and mask loss is binary cross entropy loss.

For consistency training with both datasets, we leverage the semi-supervised mechanism that uses a labeled set predicted mask as ground truth to teach unlabeled dataset pseudo masks. The consistency loss is cross-entropy loss, defined as:

$$L_u = \frac{1}{B_u} \sum \mathbb{1} \left( \max \left( p^\omega \right) \geq \tau \right) H \left( p^\omega, p^s \right), \qquad (13)$$

where $B_u$ denotes the batch size of unlabeled set, $\tau$ is a pixel-level threshold, and $p$ is pixel-level predicted probability. The overall objective function is a combination of supervised loss $L_{MaskRCNN}$ and unsupervised loss $L_u$ as:

$$L = L_{MaskRCNN} + \lambda L_u \qquad (14)$$

where $\lambda$, the relative weight of unsupervised loss, is a scalar hyperparameter.

## 3.3. Sparse-Stacking Consistency with Self-Training Scheme

Based on our earlier research, which assumed that the SBS-Stacking method worked well enough, we discovered that tuning $k$-slice for training is not always stable and produces decent results. The reason for this is that the model is not consistent enough as a result of some samples lacking or sparsely distributed slices. We further employed the mechanism that post-selects certain samples with only a few slices yet is trustworthy enough for supervised training after stacking, called the Sparse-Stacking Consistency Self-training framework. Our goal is to develop a selective re-training technique that identifies the trustworthy stacked samples so that lower-slice samples can be safely exploited.

Previous works estimate the uncertainty or reliability of an image or pixel from various angles: for example, they train two differently initialized models to predict the same unlabeled sample and reweight the uncertainty-aware loss with their disagreements [17]; they also take the final softmax output as the confidence distribution and filter low-confidence pixels by pre-defined threshold [47, 63]. We want to test the dependability with a single training model in our proposed self-training framework, eliminating the need to manually select the confidence threshold. Additionally, we eliminate untrustworthy samples using image-level information as opposed to the often-used pixel-level information for a more steady reliability rating. During training, the model is also able to pick up more comprehensive contextual patterns because of the image-level selection.

In our work, mean IOU is employed to measure the stability and reliability among pseudo-masks from the student model and predicted masks from the teacher model. Given that, with any $n \in [0; k-1]$, considering stacked image
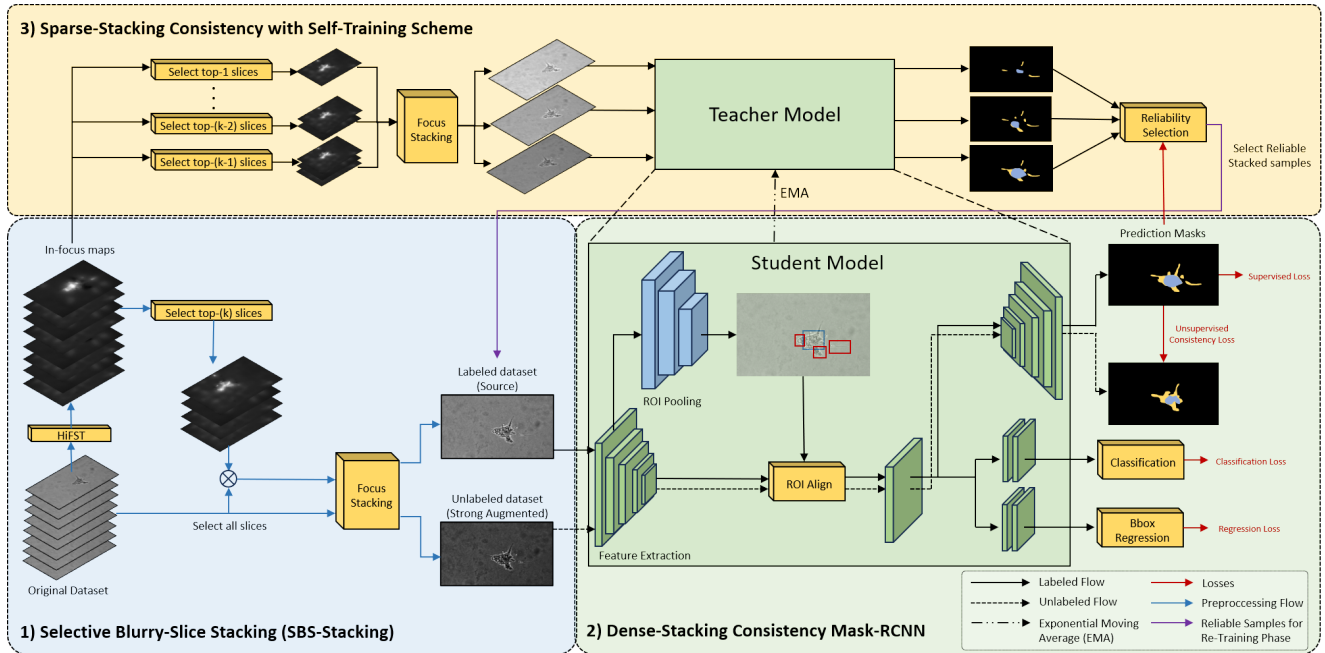
Figure 2. Proposed pipeline in detail with two main parts. 1) SBS-Stacking takes the original dataset to create partially stacked images and full focus-stacked images. 2) DSCMask R-CNN architecture that utilizes Mask R-CNN and mechanism of consistency training by treating SBS-Stacking images as a weak-augmented set and Focus-Stack images as a strong-augmented set. 3) Sparse-stacking consistency with a Self-training scheme will filter and select reliable images for the re-training stage by using a teacher model.

$u_i^n \in D^{u^n}$, The meanIOU can serve as a measurement for stability and further reflect the reliability of the unlabeled image along with the pseudo mask:

$$s_i^n = \sum_{j=1}^{K-1} meanIOU \left( M_{ij}^n, M_{iK}^n \right)$$

whereas $s_i^n$ is the stability value of an n-stacked image, showing the reliability of $u_i^n$. $K$ is the checkpoint during the training stage of DSCMask-RCNN. $M$ is the pseudo-mask predicted from the teacher and student model. After calculating each unlabeled image's stability score, we use these scores to order the entire n-slice stacked collection and compare it with the pre-defined threshold. In our experiments, the threshold $\tau$ for determining reliable sample shown as:

$$r_i = \left\{ u_i^n \,|\, s_i^n > \tau \right\}$$

We set $\tau = 0.8$ for all of our experiments.

## 4. Experiments and Results

### 4.1. Dataset

Over the course of four days, the breast cancer cell line MDA-MB-231 was used to generate spheroids in a microwell dish. Following that, the spheroids were harvested and embedded in a collagen type I gel to allow for a 3D invasion assay. Continuous monitoring was accomplished using differential interference contrast (DIC) microscopy, which captured time-lapse images at 15-minute intervals over a 28-hour period using an objective lens with a magnification factor of 20X. A series of 15 z-stack images were acquired at each time point, spanning from the bottom to the top of the spheroid invasive region, with a consistent inter-plane distance of 20 micrometers. The collected images were then analyzed further to determine the invasive dynamics and behavior of the spheroid. In total, after stacking, three experienced annotators labeled 426 stacked images from 6390 slices with three classes: Invasive area, Spheroid core, and Single Cells.

### 4.2. Implementation Details

Our experimental settings are based on PyTorch using the NVIDIA RTX 2080Ti 11GB VRAM. For the preprocessing stage, we cleaned some bad samples by hand, and then SBS-Stacking was performed to stack $k$ images. We choose $k = 8$, which is nearly half of the total 15 slices for every batch. For the training stage, our DSCMask R-CNN is built on top of the Mmdetection library. It was trained with 100 epochs, using an Adam optimizer with an initial learning rate of 0.001, weight decay of 1e-4, and a decay rate of 0.1

to minimize the total loss. Basic geometric augmentation is brought to reduce the model's bias, including rotation, flipping, and random cropping. We set fixed $\lambda = 0.1$ to all experiments for consistency training with an unlabeled set.
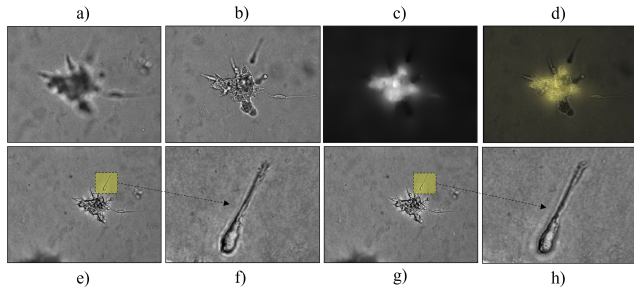
## 4.3. Qualitative Result: SBS-Stacking



Figure 3. Visualization of a) Out-of-focus slice ($Z$-13); b) Partially in-focus slice ($Z$-6); c) and d) In-focus detection map of image b) and its overlay image; e) and f): Output images of SBS-Stacking; g) and h): Output images of normal Focus-Stack.

Based on Figure 3, it can be seen that some of the slices in the cell image are out-of-focus and full of blurred parts, as image 3a. Meanwhile, image 3b contains more sharp areas but still has many blurry areas. Through the HiFST algorithm, image 3c shows its in-focus area with a grey area, which overlaps nearly 80% of the cells. In the stacked images, our proposed SBS-Stacking shows its potential with the sharp contours and less effect of blurry slices, as in image 3e and its zoom-in image 3f. The traditional Focus-Stack without cleaning blurry images shows the noisy pixels with no clearly sharp boundary of cells, as in the image 3g and image 3h.

## 4.4. Quantitative Result: DSCMask R-CNN

| | Backbone | Bounding box evaluation | | Mask evaluation | |
|---|---|---|---|---|---|
| | | mAP@[0.5:0.95] | mAP@0.5 | mAP@[0.5, 0.95] | mAP@0.5 |
| Ngo TKN [15] | R50 | - | - | 0.559 | 0.832 |
| | R101 | - | - | 0.571 | 0.846 |
| | X101-32x4d | - | - | 0.579 | 0.852 |
| | X101-64x4d | - | - | 0.586 | 0.868 |
| Mask R-CNN | R50 - FPN | 0.690 | 0.930 | 0.620 | 0.918 |
| | R101 - FPN | 0.702 | 0.946 | 0.631 | 0.932 |
| | X101-32x4d | 0.716 | 0.958 | 0.642 | 0.947 |
| | X101-64x4d | 0.724 | 0.962 | 0.650 | 0.959 |
| DSCMask R-CNN | R50 - FPN | 0.692 | 0.938 | 0.628 | 0.931 |
| | R101 - FPN | 0.704 | 0.944 | 0.636 | 0.940 |
| | X101-32x4d | 0.722 | 0.960 | **0.654** | 0.964 |
| | X101-64x4d | **0.730** | **0.966** | 0.652 | **0.967** |

Table 1. Quantitative results of Mask R-CNN and DSCMask R-CNN with various backbones compared to previous related work on stacked 3D cells dataset.

In this experiment, we compared our methods with previous related work. Similar to our current work, Ngo TKN et.al. [15] work focuses on 2D modality breast cancer cells. The work is based on the Unet++ semantic segmentation network that does not perform bounding box prediction as our two-stage approach. For a fair comparison, we implemented their methods under similar backbones. Their works proposed multiple models for performing segmentation on invasive areas, single cells, and core areas. However, we only used the invasive areas model for comparison with our methods. The hyperparameter config was set similarly to our proposed architecture.

According to Table 1, the DSCMask R-CNN shows the overall improvement on all feature extractors. In our work, we use mAP as the main evaluation metric for both instance segmentation and detection. It can be seen that Mask R-CNN shows the highest performance on the setting of X101-64x4d with 72.4% and 65.0% with $mAP@[0.5, 0.95]$. Our proposed DSCMask R-CNN utilizes the mechanism of consistency training achieved on both tasks, with 73.0% and 65.2% on $mAP@[0.5, 0.95]$ respectively. For $mAP@0.5$, our best result achieved on both tasks is up to 96.6% and 96.7%, respectively. As a consequence, the stacking method for preprocessing might not always be ideal for all samples. Consistency training can reduce the significant amount of unclear boundary problems by leveraging weak-to-strong self-training. Compared to previous work, our method shows promising results against different approaches that use semantic segmentation with the U-net family. With similar backbones, DSCMask R-CNN architecture outperforms nearly 10% with $mAP@0.5$ metric over previous work[15].

## 4.5. Ablation Studies on Proposed Framework

We compared multiple settings of our framework as ablation studies as shown in Table 2. We used mAP@[0.5,0.95] for our evaluation metric for mask prediction. It can be seen that the baseline model that uses the SBS-Stacking method and the Mask R-CNN model achieved 62.32%, lower than 1,86% compared to the setting that applied SSC Self-training. We changed Mask R-CNN into DSCMask R-CNN without SSC Self-training, which reached 63.92%, higher than using Mask R-CNN 1.6%. The combination of all methods reaches the peak at 66.52%, which is our current state-of-the-art on the differential interference contrast 3D Breast Cancer Spheroid dataset.

## 5. Conclusion

Overall, our work aims to propose a new method to handle the 3D breast cancer cells DIC microscopy dataset by stacking them with SBS-Stacking. The algorithm eliminates the unclear boundary caused by blurry slices from the original batch of images. Besides, our combination Mask

Table 2. Ablation studies of our three proposed modules using ResNet-50 RPN as backbone on 3D breast cancer spheroid dataset

| | SBS-Stacking | Mask R-CNN | DSCMask R-CNN | SSC Self-training | mAP@[0.5, 0.95] |
|---|:---:|:---:|:---:|:---:|:---:|
| Baseline | ✓ | ✓ | | | 62.32 |
| Proposed Methods | ✓ | ✓ | | ✓ | 64.18 |
| | ✓ | | ✓ | | 63.92 |
| | ✓ | | ✓ | ✓ | **66.52** |

R-CNN with weak-to-strong consistency training, named DSCMask R-CNN, also helps the model avoid biases and boosts the performance of architecture under various backbones. Lastly, our proposed self-training framework has boosted the performance by selectively utilizing the fewer-slice image for the re-training stage. Our dense-stacked and sparse-stacked handling was approached in different ways using self-training and consistency regularization, respectively. In future work, we aim to expand from instance segmentation to tracking tasks. Biological analysis will be conducted from our robust framework.

## 6. Compliance with ethical standards

The research data was conducted by the IMBSL Lab, Department of Biomedical Engineering, National Cheng Kung University. This study did not require any ethical approval.

## References

[1] S. Alireza Golestaneh and Lina J. Karam. Spatially-varying blur detection based on multiscale fused and sorted transform coefficients of gradient magnitudes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020. 2

[3] Veysel Aslantas and Ahmet Nusret Toprak. A pixel based multi-focus image fusion method. *Optics Communications*, 332:350–358, 2014. 2

[4] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 810–818. Springer, 2019. 2

[5] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6912–6920, 2021. 2

[6] Liliana R Cisneros Castillo, Andrei-Dumitru Oancea, Christian Stüllein, and Anne Régnier-Vigouroux. A novel computer-assisted approach to evaluate multicellular tumor spheroid invasion assay. *Scientific reports*, 6(1):35099, 2016. 1

[7] Martin Ducker, Valerie Millar, Daniel Ebner, and Francis G Szele. A semi-automated and scalable 3d spheroid assay to study neuroblast migration. *Stem Cell Reports*, 15(3):789–802, 2020. 1

[8] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. 2019. 2

[9] Istvan Grexa, Akos Diosdi, Maria Harmati, Andras Kriston, Nikita Moshkov, Krisztina Buzas, Vilja Pietiäinen, Krisztian Koos, and Peter Horvath. Spheroidpicker for automated 3d cell culture manipulation using deep learning. *Scientific Reports*, 11(1):14813, 2021. 1

[10] Liang Kou, Liguo Zhang, Kejia Zhang, Jianguo Sun, Qilong Han, and Zilong Jin. A multi-focus image fusion method via region mosaicking on laplacian pyramids. *PloS one*, 13(5): e0191085, 2018. 2

[11] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, page 896. Atlanta, 2013. 2

[12] Jinjiang Li, Genji Yuan, Hui Fan, et al. Multifocus image fusion using wavelet-domain-based deep cnn. *Computational intelligence and neuroscience*, 2019, 2019. 2

[13] Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8801–8809, 2021. 2

[14] Ba Hung Ngo, Ba Thinh Lam, Thanh Huy Nguyen, Quang Vinh Dinh, and Tae Jong Choi. Dual dynamic consistency regularization for semi-supervised domain adaptation. *IEEE Access*, 2024. 2

[15] Thi Kim Ngan Ngo, Sze Jue Yang, Bin-Hsu Mao, Thi Kim Mai Nguyen, Qi Ding Ng, Yao-Lung Kuo, Jui-Hung Tsai, Shier Nee Saw, and Ting-Yuan Tu. A deep learning-based pipeline for analyzing the influences of interfacial mechanochemical microenvironments on spheroid invasion

using differential interference contrast microscopic images. *Materials Today Bio*, 23:100820, 2023. 1, 6

[16] Huy T Nguyen, Thinh B Lam, Quan TD Tran, Minh T Nguyen, Dat T Chung, and Vinh Q Dinh. In-context cross-density adaptation on noisy mammogram abnormalities detection. In *2023 IEEE Statistical Signal Processing Workshop (SSP)*, pages 383–387. IEEE, 2023. 1

[17] Thanh-Huy Nguyen, Quang Hien Kha, Thai Ngoc Toan Truong, Ba Thinh Lam, Ba Hung Ngo, Quang Vinh Dinh, and Nguyen Quoc Khanh Le. Towards robust natural-looking mammography lesion synthesis on ipsilateral dual-views breast cancer analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2564–2573, 2023. 1

[18] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12674–12684, 2020. 2

[19] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018. 2

[20] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE international conference on computer vision*, pages 4714–4722, 2017. 2

[21] Vidas Raudonis, Agne Paulauskaite-Taraseviciene, and Kristina Sutiene. Fast multi-focus fusion based on deep learning for early-stage embryo image enhancement. *Sensors*, 21(3):863, 2021. 4

[22] Madhu S Sigdel, Madhav Sigdel, Semih Dinç, Imren Dinc, Marc L Pusey, and Ramazan S Aygün. Focusall: Focal stacking of microscopic images using modified harris corner response measure. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(2):326–340, 2015. 2

[23] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 2

[24] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 2

[25] Han Tang, Bin Xiao, Weisheng Li, and Guoyin Wang. Pixel convolutional neural network for multi-focus image fusion. *Information Sciences*, 433:125–141, 2018. 2

[26] Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42:245–284, 2015. 2

[27] Toan TN Truong, Huy T Nguyen, Thinh B Lam, Duy VM Nguyen, and Phuc H Nguyen. Delving into ipsilateral mammogram assessment under multi-view network. In *International Workshop on Machine Learning in Medical Imaging*, pages 367–376. Springer, 2023. 1

[28] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4248–4257, 2022. 2

[29] Louis-Bastien Weiswald, Dominique Bellet, and Virginie Dangles-Marie. Spherical cancer models in tumor biology. *Neoplasia*, 17(1):1–15, 2015. 1

[30] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4268–4277, 2022. 2

[31] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196, 1995. 2

[32] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8229–8238, 2021. 2

[33] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 2

[34] Noam E Ziv and Jackie Schiller. Differential interference contrast (dic) imaging of living cells. *CSH protocols*, 2007: pdb–prot4787, 2007. 1

[35] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 2