CellGenie: An end-to-end Pipeline for Synthetic Cellular Data Generation and Segmentation: A Use Case for Cell Segmentation in Microscopic Images

Nabeel Khalid[1,2][0000−0001−9274−3757], Mohammadmahdi Koochali[1][0000−0001−8780−253X], Duway Nicolas Lesmes Leon[1,2][0009−0007−4677−7105], Maria Caroprese[5][0009−0009−2170−1459], Gillian Lovell[4][0009−0004−5180−9704], Daniel A Porto[6][0000−0002−1021−2467], Johan Trygg[3,7][0000−0002−4239−6520], Andreas Dengel[1,2][0000−0002−6100−8255], and Sheraz Ahmed[1][0000−0002−4239−6520]

[1] German Research Center for Artificial Intelligence (DFKI) GmbH, Kaiserslautern 67663, Germany
firstname.lastname@dfki.de
[2] RPTU Kaiserslautern–Landau, Kaiserslautern 67663, Germany
[3] Sartorius Corporate Research, Sweden
[4] Sartorius, BioAnalytics, Royston, United Kingdom
[5] Sartorius, Digital Solutions, Royston, United Kingdom
[6] Sartorius, BioAnalytics, Ann Arbor, United States
firstname.lastname@sartorius.com
[7] Computational Life Science Cluster (CLiC), Umeå University, Sweden

Abstract. Cellular imaging plays a pivotal role in understanding various biological processes and diseases, making accurate cell segmentation indispensable for many biomedical applications. However, traditional methods for cell segmentation often rely on manual annotation, which is labor-intensive and time-consuming. Deep learning-based approaches for cell segmentation have shown promising results, but they require a vast amount of annotated data for training. In this context, this study presents CellGenie, an end-to-end pipeline designed to address the challenge of data scarcity in deep learning-based cell segmentation. This research proposes an innovative approach for automatic synthetic data generation tailored for microscopic image analysis. Leveraging the rich information provided by the LIVECell dataset, CellGenie generates synthetic microscopic images along with their corresponding segmentation masks for individual cells. By seamlessly integrating this synthetic data into the training process, this study enhances the performance of cell segmentation models beyond the limitations of existing annotated dataset. Furthermore, extensive experimentations are conducted to evaluate the efficacy of the generated data across various experimental scenarios. The results demonstrate the substantial impact of synthetic data generation in improving the robustness and generalization of cell segmentation models.

Keywords: cell segmentation · synthetic data · microscopic imaging · deep learning.

1 Introduction

In microscopic image analysis, cells serve as fundamental units of life and are essential for understanding various biological processes and diseases. By providing insights into cellular morphology, disease biomarkers, and drug responses, microscopic analysis facilitates advancements in precision medicine, personalized therapeutics, and innovative healthcare solutions. The foundational step in studying microscopic images involves cell segmentation, an intricate process requiring the delineation of each cell's boundary. By accurately delineating individual cells within microscopic images, segmentation facilitates disease diagnosis and treatment by identifying aberrant cellular phenotypes indicative of pathological conditions. In drug discovery, cell segmentation plays a pivotal role in screening potential therapeutic compounds, assessing their efficacy, and elucidating underlying mechanisms of action. Deep learning-based approaches for cell segmentation demand extensive volumes of fully annotated data for training, where each cell's boundary is carefully delineated. This annotation process is not only time-consuming but also very expensive [9].

In the domain of microscopic image analysis, the LIVECell dataset [2] stands out as one of the largest and most comprehensive resources in cell biology research. With over 1.6 million cells, it boasts an average cell density per image surpassing that of any other publicly available dataset in the field, reaching 313 cells—an approximately 55-fold increase compared to the EVICAN dataset [13]. Annotating cells within microscopic images presents unique challenges compared to annotating objects in natural images due to their smaller scale, higher complexity, greater variability, and increased noise. In environments where cell cultures, such as BV2, are densely packed, and in the presence of morphologically complex cell types like SH-SY5Y with their asymmetric and concave shapes, traditional methods of manually annotating cell boundaries become highly challenging. These complexities often lead to difficulties in accurately segmenting cells, compounded by the sheer volume of cells present. Moreover, the average annotation time per cell, which stands at 46 seconds within the LIVECell dataset, underscores the labor-intensive nature of the task. To address these challenges, this study introduces CellGenie, a synthetic microscopic data generation approach leveraging the LIVECell dataset to automatically generate microscopic images with cell masks. Fig. 1 showcases the comparison between real and generated images from cell cultures A172 and BV-2, where each cell is delineated by a yellow boundary mask.

The finding of this study underscores the substantial benefits of synthetic data generation in enhancing the robustness and generalization of cell segmentation models. By presenting CellGenie as an accessible and cost-effective solution, this research contributes to the advancement of biomedical imaging and computa-

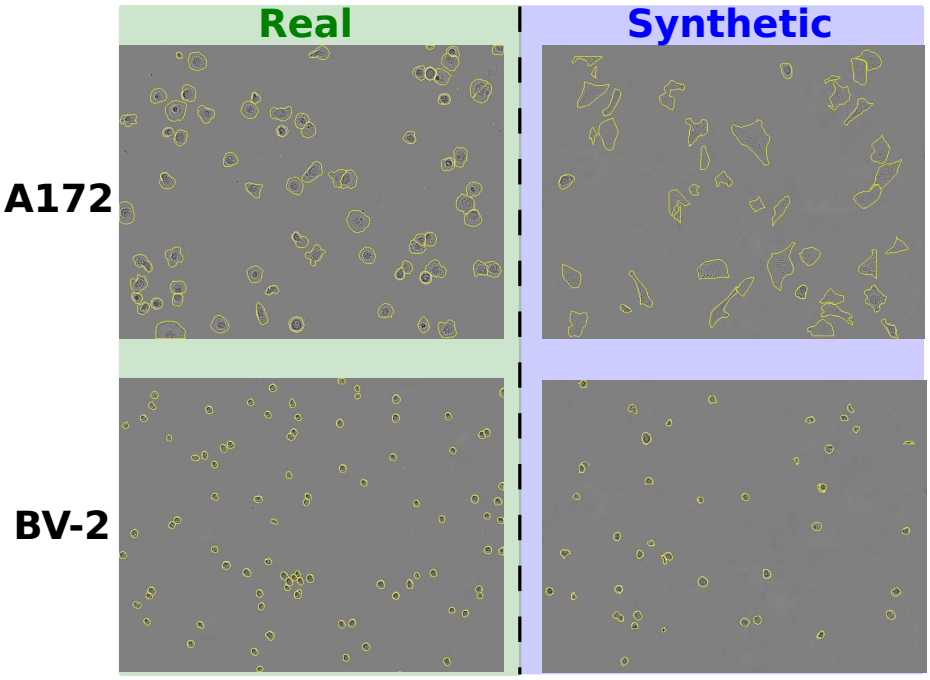

Fig. 1: Comparison of real and synthetic images from cell cultures A172 and BV-2. Each cell is delineated by a yellow boundary mask.

tional biology, opening new avenues for scalable and accurate cellular analysis. The main contributions of this study are as follows:

1. This study introduces CellGenie, an end-to-end pipeline for synthetic cellular data generation, leveraging the LIVECell [2] dataset to produce microscopic images for cell segmentation. CellGenie also automates the generation of segmentation masks for cells within the images.
2. By integrating this synthetic data into the training process, this study enhances the performance of cell segmentation models beyond the limitations of the existing annotated LIVECell dataset.
3. Extensive experimentation is undertaken to assess the effectiveness of the generated data under diverse experimental conditions.

2 Related Work

In the past decade, there has been a remarkable advancement in deep learning-based cell analysis, notably with the introduction of the U-net architecture by Ronneberger et al. [12] in 2015. Despite being trained on only 35 images, the U-net model surpassed all competitors in the 2015 ISBI cell tracking and segmentation challenge. This success catalyzed a series of significant developments

in image-based cellular research, leading to the creation of pioneering algorithms such as CellPose [14], DeepCeNS [7], DeepCIS [8], and DeepMuCS [6]. However, obtaining the annotations necessary for training deep learning models is often a laborious and challenging task. To mitigate this challenge, researchers have proposed weakly supervised or semi-supervised learning approaches to alleviate the annotation burden. Weakly supervised techniques, such as points [9, 5] and missing annotations [3] have been proposed. Khalid et al. (2023) [5] introduced a method for weakly supervised cell segmentation, leveraging multiple points along with a bounding box for each cell. Their approach achieved 99.8% of the performance attained through fully supervised methods, using 8-point labels and bounding boxes. Importantly, this approach substantially reduced the time needed for data annotation, being 3.24 times faster than annotating the full mask.

While weakly supervised and semi-supervised learning approaches have provided significant relief from the laborious task of manual annotation in deep learning-based cell analysis, they still entail considerable time and effort. Despite their ability to reduce annotation burdens through techniques like image tags, points, and missing annotations, these approaches necessitate expert knowledge for accurate labeling. For instance, while Khalid et al.'s method achieved remarkable performance with minimal annotation requirements, it still demands expertise to select appropriate points and bounding boxes for each cell. Thus, while these approaches offer notable efficiency gains compared to fully supervised methods, they remain time-consuming and reliant on expert input.

3 CellGenie-Generation: Synthetic Cellular Data Generation

Algorithm 1 illustrates the working of the proposed approach for synthetic cellular data generation. The proposed pipeline can be divided into four modules:

3.1 Cell Extraction from Original Images

In the initial phase, the cells are systematically extracted from various cultures included in the training set of the LIVECell dataset. Utilizing the segmentation mask corresponding to each cell, this careful extraction process guarantees the accuracy of cellular features. Additionally, the cells extracted are classified by their cell type. The background images are also extracted in this phase. This step lays the groundwork for the synthetic image generation framework, offering a broad and varied range of cell types crucial for subsequent analyses.

3.2 Random Selection within Normal Ranges

Each culture within the LIVECell dataset exhibits unique cell characteristics; for instance, BV-2 cell culture images may contain as many as 3000 cells per image, while Huh7 cell cultures typically have a maximum of 100 cells per image.

Algorithm 1 CellGenie Synthetic Data Generation

Input: LIVECellDataset
procedure *Extract_Cells*(LIVECellDataset) ▷ **Module 1**
    ExtractedCells ← empty dictionary
    for each Image in LIVECellDataset **do**
        Get Culture from Image
        for each Cell in Image **do**
            Get Area from Cell
            ExtractedCells[Cell] ← (Area, Culture)
        end for
    end for
    return ExtractedCells
end procedure
procedure *Select_Cells*(CellCulture, ExtractedCells) ▷ **Module 2**
    Get AreaRange, PopulationRange from CellCulture
    N ← Random integer in PopulationRange
    CellCandidates ← empty set
    for each Cell in ExtractedCells **do**
        if Cell(Area) in AreaRange **and** Cell(Culture) = CellCulture **then**
            add Cell to CellCandidates
        end if
    end for
    SelectedCells ← random subset of size N from CellCandidates
    return SelectedCells
end procedure
procedure *Generate_Image*(CellCulture, SelectedCells) ▷ **Module 3**
    NewImage ← SyntheticBackground
    for each Cell in SelectedCells **do**
        NewImage ← randomly locate Cell
        if there is overlap resolution **then**
            Handle Cell to overlap
            Annotate Cell in AnnotationCOCO
        end if
    end for
    return NewImage, AnnotationCOCO
end procedure
procedure *Create_Dataset*(LIVECellDataset, CellCultures, ImagesPerCulture)
    ExtractedCells ← *Extract_Cells*(LIVECellDataset) ▷ **Module 4**
    NewDataset ← empty set, NewAnnotation ← empty set
    for CellCulture in CellCultures **do**
        for $n = 1, \ldots,$ ImagesPerCulture **do**
            SelectedCells ← *Select_Cells*(CellCulture, ExtractedCells)
            NewImage, AnnotationCOCO ← *Generate_Image*(CellCulture, SelectedCells)
            Add NewImage to NewDataset
            Add AnnotationCOCO to NewAnnotation
        end for
    end for
    return NewDataset, NewAnnotation
end procedure

Furthermore, each cell culture displays specific area ranges characteristic of its cell distribution. By adhering to the typical cell count and area ranges for each culture, the proposed process ensures the biological authenticity of the synthetic cultures. This approach mirrors the natural variability observed in real-world cell populations within a controlled experimental framework.

3.3 Randomized Distribution on Synthetic Backgrounds

This step involves the placement of cells onto synthetic backgrounds and is further divided into two key sub-steps:

Initial Placement Cells are randomly positioned across the background, and their new coordinates are carefully recorded in the COCO format, facilitating integration with existing bioinformatics tools and datasets.

Segmentation and Overlap Resolution This step addresses instances of overlap resulting from the random placement of cells. A corrective segmentation procedure is implemented to create new segmentation masks for partially obscured cells, ensuring an accurate representation of each cell's visible portion.

3.4 Synthesis of the Final Dataset

After completing the distribution and adjustment phases, the process is iterated for a specified number of images. Through this iterative process, we generate a large dataset consisting of synthetic cell culture images. Each image in this dataset represents various realistic scenarios, enhancing the dataset's robustness and versatility for further analysis and experimentation.

4 CellGenie-Segmentation: Cell Segmentation Pipeline

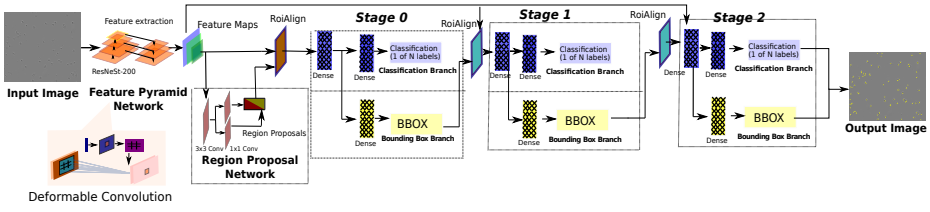

Fig. 2: **System overview of the CellGenie-Segmentation pipeline for cell segmentation.** The input image is passed to the proposed pipeline and the output image with a segmentation mask for each cell is produced.

Fig. 2 illustrates the system overview of CellGenie-Segmentation. The proposed method is based on Cascade Mask R-CNN [1], Feature Pyramid Network [11], ResNeSt-200 [16] and Deformable Convolution. The proposed pipeline is composed of three main modules: Feature Extraction, Groundtruth Association, and Prediction Head.

4.1 Feature Extraction

The purpose of this module in the proposed method is to extract feature maps from the input image at different scales. The feature extraction module is composed of Feature Pyramid Network (FPN) [10] along with ResNeSt-200 [16]. FPN combines the low resolution, semantically strong features with high-resolution, semantically weak features. It consists of a bottom-up pathway and a top-down pathway. The bottom-up pathway extracts feature maps from the input image at different scales using a series of convolutional layers. ResNeSt-200 with deformable convolution is used as a feed-forward CNN architecture in the bottom-up pathway of the proposed approach. The top-down pathway merges features from the bottom-up pathway using lateral connections and upsampling with features from higher-resolution layers to create a feature pyramid.

4.2 Groundtruth Association

The multi-scale features from the Feature Extraction module are passed onto the Groundtruth Association module. Here, the Region Proposal Network (RPN) detects the regions that contain cells and matches them to the groundtruth. Matching is achieved by generating anchors on the input image, which are then matched to the ground truth based on the Intersection over Union (IoU) computation between the anchors and ground truth. If IoU is larger than the defined threshold of 0.7, the anchor is linked to one of the groundtruth boxes and assigned to the foreground. If the IoU is greater than 0.3 and smaller than 0.7, it is considered background and otherwise ignored. At the final stage of RPN, we choose 3,000 region proposal boxes from the predicted boxes.

4.3 Prediction Head

At the prediction head, we have groundtruth boxes, proposal boxes from RPN, and feature maps from FPN. The job of the prediction head is to predict the class, bounding box, and binary mask for each region of interest. A 3-stage Cascade Mask R-CNN [1] is used as the prediction head, which is an extension of Mask R-CNN [4] with the addition of cascade stages to further improve the segmentation performance. The Cascade Mask R-CNN enhances segmentation performance by introducing cascade stages with increasing Intersection over Union (IoU) thresholds (0.5, 0.6, and 0.7) to refine predictions. A mask branch is added in the final stage parallel to the box branch, which is composed of a small Fully Convolutional Network (FCN) to predict a segmentation mask for each RoI in a pixel-to-pixel manner to achieve the task of instance segmentation.

5 Dataset

In the field of cell biology research, publicly available datasets play a crucial role in advancing the understanding of cellular processes. Among these, the LIVECell dataset, as described by Edlund et al. (2021) [2], stands out for its vast size and high quality. With over 1.6 million cells spread across 5,239 images, the LIVECell dataset is one of the most extensive and comprehensive resources available for cell biology studies. Notably, it encompasses eight distinct morphological cell cultures, providing researchers with a diverse array of cellular structures to analyze. A notable characteristic of the LIVECell dataset is its exceptionally high cell density, averaging 313 cells per image. This density far exceeds that of other datasets like EVICAN (Schwendy et al., 2020) [13], making LIVECell a rich source of data for cellular analysis. Despite its complexity, the dataset's high cell density presents researchers with a valuable opportunity to study densely populated cellular environments, which are often encountered in real-world scenarios.

For the purpose of this research, the original LIVECell train set is called LIVECell_Base, and in addition to that, two subsets called LIVECell+800 and LIVECell+1600 are generated. The LIVECell+800 and LIVECell add 800 and 1600 more images (100 and 200 images per cell culture, respectively) to the original LIVECell train set, respectively. The same validation and test set is used for the training and evaluation. Table 1 gives more insights into the total number of images in each subset and the total number of cell instances for each setting.

Table 1: Summary statistics of images and cells in different subsets for training.

LIVECell_Base		**LIVECell+800**		**LIVECell+1600**		**LIVECell_val**		**LIVECell_test**	
Images	**Cells**	**Images**	**Cells**	**Images**	**Cells**	**Images**	**Cells**	**Images**	**Cells**
3253	1018576	4053	1131335	4853	1272461	570	181609	1564	462261

6 Evaluation Metrics

Standard COCO evaluation protocol [11] is adapted to evaluate the performance of the proposed synthetic cellular data generation approach with the same modification of the area ranges and the maximum number of detections as reported in [2]. For the evaluation, the mean average precision for both object detection and segmentation tasks at different IoU thresholds of 0.5 (mAP50), 0.75 (mAP75), and 0.5:0.95 in the steps of 0.05 (mAP) is reported. To identify the performance of the model on objects of varied sizes, we have also included mAP for different area ranges.

7 Experimental Setup

The impact of synthetic data on microscopic image analysis is investigated through a series of experiments, each presenting a distinct scenario. In the first experimental setting, termed "LIVECell Base vs. Synthetic Enhancement," two subsets of synthetic data (800 and 1600 images) are incorporated into the LIVECell Base for training. The performance of cell detection and segmentation is then compared to that of the model trained solely on the LIVECell_Base. Moving to the second setting, "Individual Cell Culture Analysis: Base vs. Enriched," each cell culture within the LIVECell dataset is trained independently, with 100 and 200 synthetic images added respectively to enrich the dataset. This allows for performance evaluation against the baseline of each cell culture. Lastly, in the third experimental setting, "Subset Analysis: LIVECell vs. Synthetic Enrichment," four distinct subsets are extracted from the original LIVECell dataset. To each subset, 800 and 1600 images are appended separately, and their performance is compared with the baseline subset.
The training for all experimental settings was conducted using eight NVIDIA V-100 GPUs. Transfer learning was employed to train CellGenie, utilizing the MS-COCO pre-trained model [11] for all settings. The pre-trained model underwent fine-tuning using Stochastic Gradient Descent (SGD) [15]. Throughout the experiments, a base learning rate of 0.02 and a momentum of 0.9 were maintained. Additionally, anchor sizes and aspect ratios were set uniformly across all settings, with sizes (8,16,32,64,128) and ratios (0.5, 1, 2, 3, 4) utilized. The selection of checkpoints for evaluation was based on validation average precision.

7.1 Experimental Setting 1: LIVECell Base vs. Synthetic Enhancement

In this experimental setting, the effects of integrating two distinct sets of synthetic images, comprising 800 and 1600 images respectively, into the LIVECell training dataset were explored. The primary objective was to investigate how the inclusion of synthetic data could enhance the performance of cell detection and segmentation algorithms.
Table 2 provides a comprehensive overview of the performance of the proposed pipeline, CellGenie, on the generated synthetic data. The analysis revealed a notable improvement in segmentation Average Precision (AP) of 0.3% when incorporating just 800 synthetic images alongside the 3253 original images in the LIVECell training set. Additionally, an extra 0.2% enhancement in cell detection and segmentation performance was observed with the inclusion of 1600 synthetic images.

7.2 Experimental Setting 2: Individual Cell Culture Analysis: Base vs. Enriched

This experimental setting explores the effects of enriching each cell culture dataset separately with additional data. Specifically, 100 and 200 images per cell

Table 2: Performance comparison of CellGenie on LIVECell training data with and without integration of synthetic images. Results are reported on the LIVECell test set, with the best and second-best performances highlighted in green and blue colors, respectively.

Train Dataset	AP		AP50		AP75		APs		APm		APl	
	Det.	Seg.	Det.	Seg.	Det.	Seg.	Det.	Seg.	Det.	Seg.	Det.	Seg.
LIVECell_Base	48.43	47.89	81.44	80.79	51.41	51.64	48.50	45.75	**49.50**	48.33	**54.18**	56.94
LIVECell+800	**48.52**	**48.21**	**81.91**	**81.37**	**51.55**	**51.90**	**48.67**	**45.93**	48.23	**48.43**	**53.99**	**57.10**
LIVECell+1600	**48.67**	**48.07**	**81.90**	**81.26**	**51.79**	**51.89**	**49.02**	**46.16**	**49.40**	**48.73**	53.05	**56.01**

culture are introduced to expand the original training dataset for each respective cell culture. This additional data is carefully generated to capture specific properties and morphological characteristics unique to each cell culture. The objective of this setting is to uncover the segmentation challenges associated with particular cell cultures and explore methods to improve the performance of these challenging segments by incorporating more synthetic data derived from the existing dataset. This approach sheds light on the intricate complexities inherent in segmenting certain cell cultures and offers insights into strategies for leveraging synthetic data to enhance segmentation performance in such cases.
Fig. 3 illustrates the results obtained for each cell culture, comparing the performance using the base cell culture training data with that achieved by incorporating an additional 100 and 200 images for each respective cell culture. For instance, in the case of cell culture A172, a notable improvement in segmentation performance of 1.4% and 1.3% is observed for the Plus100 and Plus200 models, respectively, compared to the Base model. Similarly, for BT-474, BV-2, Huh7, SH-SY5Y, SkBr3, and SK-OV-3 cell cultures, enhancements in segmentation performance ranging from 1% to 2% are achieved with the Plus100 and Plus200 trained models when compared to the Base models.

7.3 Experimental Setting 3: Subset Analysis: LIVECell vs. Synthetic Enrichment

In this experimental setting, subsets of the complete LIVECell dataset, encompassing 2%, 4%, 5%, 25%, and 50% of the total dataset, are systematically investigated to assess their impact on the complete test set. Moreover, synthetic images—800 and 1600 in total—are incorporated into each subset individually to asses their respective impacts. The primary objective of this setting is to evaluate how these subsets, varying in size, affect the overall performance on the test set when supplemented with synthetic data, as compared to the performance of the Subset_Base model.
Fig. 4 showcases the performance outcomes attained by models trained on varying percentages of the LIVECell dataset, both with and without the inclusion of synthetic data. Notably, for the 2% subset of the LIVECell dataset, we observe enhanced segmentation performance of 1.8% and 2.4% for Subset+800 and Subset+1600, respectively, in comparison to the Subset_Base model. Similarly, for

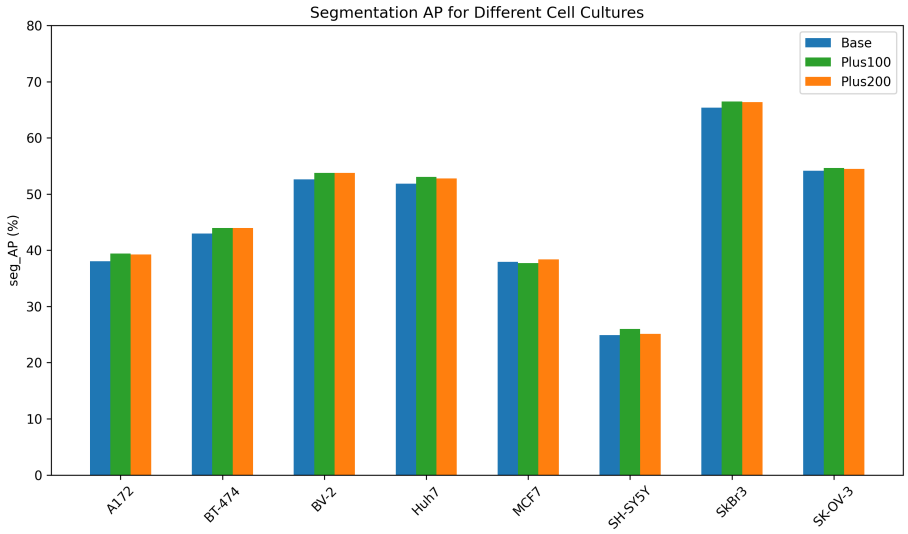

Fig. 3: Segmentation performance comparison for different cell cultures with varying numbers of additional images. Plus100 and Plus200 denote the models trained with an additional 100 and 200 images per cell culture, respectively, compared to the Base model.

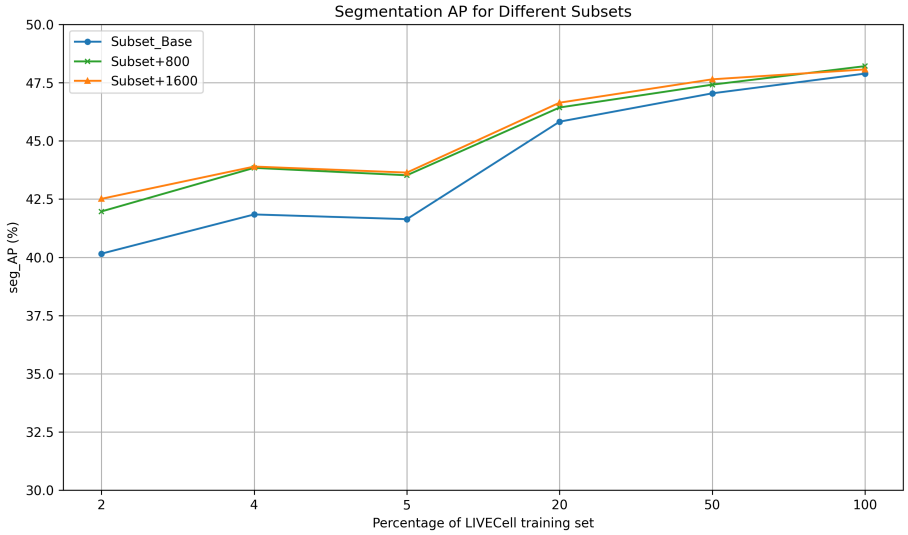

Fig. 4: Performance comparison of models trained on varying percentages of the LIVECell dataset, with and without synthetic data integration. Subset+800 and Subset+1600 denote the models trained with an additional 800 and 1600 images per subset, respectively, compared to the Subset_Base model.

the 4% and 5% subsets, the performance improvement with Subset+800 is 2% and 1.9%, respectively. Meanwhile, with Subset+1600, the performance enhancement for the same subsets is 2.1% and 2%, respectively. It's worth noting that augmenting the 2% subset results in better performance compared to using 5% of real data. Remarkably, the 50% trained model, coupled with 1600 synthetic images, achieves a segmentation performance of 47.65%, representing 99.5% of the performance attained by the model trained on 100% of the LIVECell dataset.

8 Analysis and Discussion

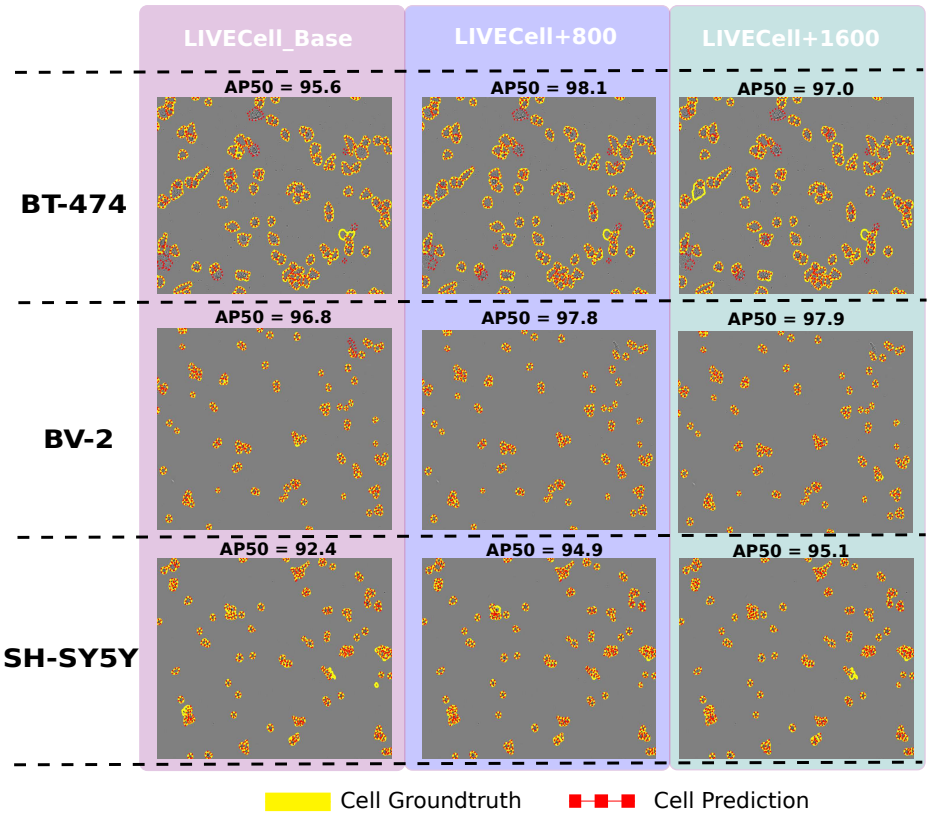

Fig. 5: Inference results showcasing segmentation performance on sample images using models trained on datasets: LIVECell_Base, LIVECell+800, and LIVECell+1600. Ground truth masks (solid yellow lines) and model predictions (dotted red lines) are depicted.

This section discusses the outcomes from the three experimental settings, focusing on their broader implications. In particular, experimental setting 1

(LIVECell Base vs. Synthetic Enhancement) evaluates synthetic data generation using the CellGenie pipeline. The findings reveal significant improvements in cell analysis algorithm accuracy and efficacy when integrating synthetic data. These results highlight the potential of synthetic data to enhance microscopic image analysis techniques, suggesting that leveraging synthetic data can address the complexities of cell analysis and lead to more robust and accurate methodologies in biomedical research. Fig. 5 shows inference results on various samples using models trained on LIVECell_Base (purple), LIVECell+800 (blue), and LIVECell+1600 (green). Solid yellow lines represent the ground truth masks, while dotted red lines depict model predictions. Each row demonstrates the models' qualitative performance on identical images from different cell cultures for comparison. Segmentation average precision scores at the IoU threshold of 0.5 (AP50) are shown atop each prediction sub-image. In the first row (BT-474 cell culture), AP50 scores are 95.6%, 98.1%, and 97.0% for LIVECell_Base, LIVECell+800, and LIVECell+1600, respectively. For BV-2 and SH-SY5Y cell cultures, the LIVECell+1600 model achieves the highest AP50 scores of 97.9% and 95.1%, respectively.

Experimental Setting 2 (Individual Cell Culture Analysis: Base vs. Enriched) evaluates segmentation performance on individual cell cultures in the LIVECell dataset, comparing results with and without additional synthetic data. The integration of synthetic data improves segmentation performance by 1% to 2% across various cell cultures, including A172, BT-474, BV-2, Huh7, SH-SY5Y, SkBr3, and SK-OV-3. These findings emphasize the importance of synthetic data in addressing unique segmentation challenges. Notably, SH-SY5Y cell culture, with its complex neuronal morphologies, shows significant improvement with additional synthetic data, highlighting the potential of synthetic data to enhance segmentation performance for diverse cell types.

In the experimental setting 3 (Subset Analysis: LIVECell vs. Synthetic Enrichment), synthetic images (Subset+800 and Subset+1600) are introduced into subsets comprising 2%, 4%, 5%, 25%, and 50% of the LIVECell training dataset, and their performance is compared to Subset_Base. Notably, by incorporating 1600 synthetic images generated using CellGenie into the 4% subset of the LIVECell training dataset—comprising only 131 images compared to the complete train set of 3253 images—we achieved 91.7% of the performance attained with the complete LIVECell train set. Similarly, the addition of 1600 synthetic images to the 50% subset of the LIVECell dataset enabled the model to achieve 99.5% of the performance attained by the model trained on 100% of the LIVECell dataset. These findings underscore the substantial performance gains achievable through the integration of synthetic data across various subset sizes, with results approaching those obtained from models trained on the entire dataset. This highlights the potential of synthetic data to bridge the performance gap between limited subset sizes and the complete dataset, offering promising avenues for efficient model training and deployment in microscopic image analysis tasks.

The proposed approach, CellGenie, for synthetic cellular data generation, has initiated a new era in the realm of microscopic image analysis. By annotating a

subset of images and subsequently generating synthetic data using the CellGenie pipeline, researchers can achieve commendable performance without the need to annotate the entire dataset. The results obtained from the aforementioned experiments serve as compelling evidence of the efficacy of this approach. Moreover, CellGenie empowers researchers to enhance the performance of models, particularly in tackling challenging cell cultures. By augmenting such datasets with additional synthetic data, the performance of segmentation models can be notably improved, as demonstrated in the experimental findings. Furthermore, CellGenie contributes to cost reduction by eliminating the need for specialized expertise and reducing annotation expenses. Additionally, the automation of the annotation process enables the rapid generation of annotated images at scale, facilitating streamlined analysis of large-scale datasets. This unprecedented scalability and speed not only accelerate research efforts but also enable researchers to explore new avenues of discovery and insight in microscopic image analysis and cell segmentation tasks.

9 Conclusion

CellGenie presents an innovative approach to synthetic cellular data generation coupled with annotation masks, offering a promising avenue for enhanced microscopic image analysis. The observed improvements in segmentation performance underscore the potential of synthetic data in refining model efficacy. Moreover, the study findings suggest significant time savings in data annotation efforts and reduced dependence on specialized expertise, democratizing access to advanced image analysis tools. The study highlights the efficacy of synthetic data integration in addressing challenges posed by complex cell cultures, leading to performance enhancements. Furthermore, this study demonstrates that by using only a small percentage of the original dataset with the addition of synthetic data, 99.5% of the complete data performance can be achieved. By leveraging synthetic datasets with annotation masks, researchers can explore new avenues of inquiry and accelerate the pace of discovery in diverse fields. Future work aims to refine the data generation pipeline by incorporating additional features such as resizing, zooming, and flipping of cells. These enhancements hold the potential to further enhance the diversity and realism of synthetic datasets, thereby fostering more robust and adaptable models for microscopic image analysis.

References

1. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)
2. Edlund, C., Jackson, T.R., Khalid, N., Bevan, N., Dale, T., Dengel, A., Ahmed, S., Trygg, J., Sjögren, R.: Livecell—a large-scale dataset for label-free live cell segmentation. Nature methods (2021)

3. Guerrero-Peña, F.A., Fernandez, P.D.M., Ren, T.I., Cunha, A.: A weakly supervised method for instance segmentation of biological cells. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer (2019)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision (2017)
5. Khalid, N., Froes, T.C., Caroprese, M., Lovell, G., Trygg, J., Dengel, A., Ahmed, S.: Pace: Point annotation-based cell segmentation for efficient microscopic image analysis. In: International Conference on Artificial Neural Networks. pp. 545–557. Springer (2023)
6. Khalid, N., Koochali, M., Rajashekar, V., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepmucs: A framework for co-culture microscopic image analysis: From generation to segmentation. In: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE (2022)
7. Khalid, N., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepcens: An end-to-end pipeline for cell and nucleus segmentation in microscopic images. In: 2021 International Joint Conference on Neural Networks (IJCNN). IEEE (2021)
8. Khalid, N., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Deepcis: An end-to-end pipeline for cell-type aware instance segmentation in microscopic images. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE (2021)
9. Khalid, N., Schmeisser, F., Koochali, M., Munir, M., Edlund, C., Jackson, T.R., Trygg, J., Sjögren, R., Dengel, A., Ahmed, S.: Point2mask: A weakly supervised approach for cell segmentation using point annotation. In: Medical Image Understanding and Analysis: 26th Annual Conference, MIUA 2022, Cambridge, UK, July 27–29, 2022, Proceedings. Springer (2022)
10. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2017)
11. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. Springer (2014)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer (2015)
13. Schwendy, M., Unger, R.E., Parekh, S.H.: Evican–a balanced dataset for algorithm development in cell and nucleus segmentation. Bioinformatics (2020)
14. Stringer, C., Wang, T., Michaelos, M., Pachitariu, M.: Cellpose: a generalist algorithm for cellular segmentation. Nature Methods (2020)
15. Wijnhoven, R.G., de With, P.: Fast training of object detection using stochastic gradient descent. In: 2010 20th International Conference on Pattern Recognition. IEEE (2010)
16. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)