Fluorescent Neuronal Cell Counting using a modified ResUnet Model with Attention Gates

Aanya Tashfeen
Stanford University
aanya@stanford.edu

Abstract

Accurate cell counting is a critical task in biological image analysis. In this paper, we evaluate several variants of the c-ResUnet architecture, including the integration of attention gates and different activation functions (ReLU and ELU), to improve segmentation performance in challenging cell imaging conditions. Performance is evaluated using precision, recall, F1 score, MAE, and MPE metrics. Our results show that while attention-based models offer improved segmentation in cluttered regions, they can also lead to overconfident predictions that merge adjacent cells.

1. Introduction

In this paper, we will investigate using a deep learning model to automate cell counting from microscopy images. Manual cell counting is time-consuming, subjective, and prone to human error, especially when dealing with densely populated or overlapping cells. It is often done to ensure experimental accuracy, consistency, and to monitor cell health and viability. Automating this process could significantly improve efficiency and reproducibility in biomedical research and clinical diagnostics, as well as potentially improving accuracy. Modern-day non-manual cell counting techniques are often found to be inaccurate and don't work for smaller cells. This task is interesting not only due to its direct real-world relevance, but also because it poses unique challenges like variable cell shapes, densities, and overlaps.

In this paper, we focus on replicating and building on a study conducted by the University of Bologna, that focuses on evaluating the effectiveness of "Automating cell counting in fluorescent microscopy through deep learning with c-ResUnet" [1]. This paper introduces c-ResUnet, a U-Net-inspired deep learning model for automatic cell counting via binary segmentation (a threshold was calculated to determine whether the cell contained a cell or not). The model segments fluorescent neuronal cells and retrieves their count by detecting these individual objects. Through ablation studies, the authors show that performance is significantly improved by using weight maps that emphasize cell boundaries, especially in images with many clumps of cells. The paper also found that c-ResUnet outperforms similar models in both detection (F1 score = 0.81) and counting accuracy (MAE = 3.09). The model and dataset are publicly released to support further research in biological imaging and deep learning applications, and so we used both the model code and the dataset for this paper [6]. In our paper, we use the same dataset and baseline c-ResUnet model to conduct a study that investigates the impact of incorporating attention gating into the architecture. Attention gates help the model focus on relevant regions of the image, by learning information such as what the target structures are shaped like and ultimately filtering out less informative features, which may improve segmentation performance in cluttered or low-contrast images [4].

One challenge in this domain is the limited availability of well-annotated microscopy data of every cell type, which makes it difficult to train large models of any type of cell of interest effectively. Additionally, we were constrained by limited computational resources and time. Specifically, we were running into memory (RAM) issues when using larger sample sizes. While batch size reduction and training-time data augmentation helped mitigate this to some extent, we were still limited by how much data we could process at once. This paper explores whether architectural modifications, such as attention gating, can help compensate for small sample sizes. Our results show that while attention-based models offer improved segmentation in cluttered regions, large sample sizes are still needed for a strong cell counting model.

1.1. Related Work

The accurate segmentation and counting of cells in microscopy images are a critical step for many biological and medical applications, yet it remains challenging due to overlapping cells, varying cell morphology, and imaging artifacts. Traditional convolutional neural network architectures like c-ResUnet have demonstrated strong performance in semantic segmentation of cells, providing pixel-wise delineation but often struggling with precise instance separation in crowded regions. On the other hand, transformer-based instance segmentation models [2] such as Cell-DETR (attention-based cell detection transformer) leverage attention mechanisms to explicitly distinguish individual cell instances, potentially offering improved accuracy in images with many clumps of cells. While researching transformer-based models like DETR that use attention mechanisms for improved instance segmentation, we became interested in whether integrating attention more directly into convolutional architectures could enhance performance on cell segmentation tasks. This led me to explore attention gating [4], a technique that incorporates spatial attention within CNNs to help the model focus on relevant image regions, and hasn't been applied to cell counting with ResUnets in previous literature. Inspired by this, I decided to investigate how adding attention gating to the c-ResUnet architecture affects segmentation and counting accuracy. This study aims to directly compare the performance of c-ResUnet and c-ResUnet with attention gating mechanisms on the same microscopy dataset to evaluate their respective strengths and limitations in cell segmentation and counting, ultimately informing model selection for automated microscopy analysis pipelines.

Recent work by the University of Austria, Innsbruck in 2024 [3] highlights the promise of vision transformers (ViTs) for weakly-supervised microorganism counting. The paper found that ResNets perform better than vision transformers on a variety of microorganism datasets mainly because they handle small datasets and low-density images more effectively. Their skip connections help train deeper networks without overfitting. Vision transformers generally need large datasets to perform well, but some advanced vision transformer variants can still compete when multi-scale features are effectively leveraged. In cell counting, multi-scale features mean detecting both tiny details like the shape of individual cells (small scale), and broader patterns like how cells are spaced or clustered across the whole image (large scale). While we don't evaluate this model, it highlights the importance of attention-based mechanisms in vision.

1.2. Dataset

The Fluorescent Neuronal Cells dataset from the University of Bologna [1] consists of 283 high-resolution microscopy images of mouse brain slices of size 1600 by 1200, where neurons are highlighted via fluorescence labeling. Due to variability in brightness, contrast, and complex cell shapes, preprocessing was necessary to facilitate accurate segmentation. The researchers applied Gaussian blurring to reduce noise and used automatic histogram-based thresholding to generate initial binary masks identifying potential neuronal cells. These masks were then manually parsed to remove false positives and artifacts, ensuring high-quality ground-truth labels for training. This combined automatic and manual pipeline helped create reliable segmentation masks despite the challenges posed by image variability, artifacts, and clustered cells, enabling robust model training and evaluation. These masks and images were cropped to 512 by 512 to create more data. Data augmentation was also applied to increase the size of the dataset, but was not used for our new proposed model architectures. They provided this dataset and the cropping method, so this was used for this paper.

2. Methods

Researchers at the University of Bologna [1] approached the task of segmenting and counting fluorescently labeled neurons using a supervised learning framework with convolutional neural networks. The input to the model is a picture of cells taken under a fluorescent microscope. Their main and best performing architecture, c-ResUNet, shown in Figure 1, includes modifications such as an initial 1×1 convolution and an additional residual block with 5×5 filters to improve context understanding, especially for overlapping cells.

In ResUnets, residual connections are implemented within each encoder and decoder block, where the input to the block is added back to its output after a series of convolutional layers. This helps mitigate vanishing gradients and enables deeper feature learning by allowing the network to learn residual mappings. Skip connections concatenate encoder and decoder layers and this ultimately makes the loss landscape more smooth and easier to train [5]. To improve the relevance of the features passed through these skip connections, we proposed applying attention gates to this c-ResUnet architecture as attention gates learn to ignore irrelevant regions and highlight important features, such as the neuronal cells, from the encoder [4]. Each attention gate takes the encoder feature map and the decoder's gating signal, and produces an attention-weighted output that is concatenated with the decoder's upsampled feature map before further processing. Figure 2 is a schematic of what

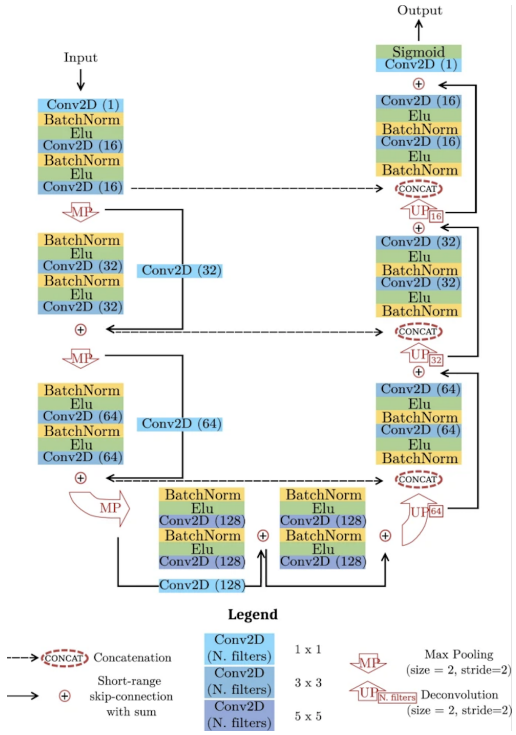

Figure 1. c-ResUnet architecture [1]

the proposed additive feature would look like. The

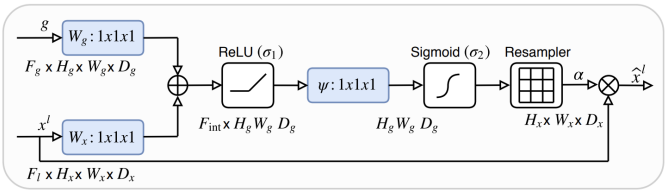

Figure 2. Attention Gate Schematic [4]

fluorescent neuronal cells dataset was split into training, validation, and test sets. Our training to validation ratio was 80% to 20%. From each image, twelve 512×512 sub-images were extracted and augmented using rotation, noise, brightness variation, and elastic deformation. Training used the Adam optimizer with early stopping and a weighted binary cross-entropy loss. This loss was used to address class imbalance in our segmentation task as there were more negatives (i.e. the background) than positive classes (the cells). So, we assign different weights to the positive (cell) and negative (background) classes. By increasing the weight for the minority class, the model is penalized more heavily for misclassifying important but underrepresented pixels. Formally, for each pixel, the loss is defined as:

$$\mathcal{L} = -w_p \cdot y \cdot \log(\hat{y}) - w_n \cdot (1 - y) \cdot \log(1 - \hat{y}) \quad (1)$$

where $y \in \{0, 1\}$ is the ground truth label, $\hat{y} \in [0, 1]$ is the predicted probability, and $w_p = 1.5, w_n = 0.5$ are the weights for the positive and negative classes, respectively, as chosen in [1]. The weights are chosen to compensate for class imbalance, encouraging the model to better learn the minority class.

We tested 4 model architectures.

Model	Activation	Optimizer	Augmentation	Sample Size	Batch Size
c-ResUnet + Attention (ELU)	ELU	Adam (10^{-4})	Yes	600	4
c-ResUnet + Attention (ReLU)	ReLU	Adam (10^{-4})	Yes	600	4
c-ResUnet (Morelli & Clissa [1])	ELU	Adam (10^{-4})	Yes	Full	8
Simple c-ResUnet (fewer samples)	ELU	Adam (10^{-4})	No	500	8

Table 1. Summary of model architectures and training configurations.

To understand the impact of architectural and training decisions, we compared four versions of c-ResUnet. One model we evaluated was the original c-ResUnet model from Morelli and Clissa [1] which used the entire 2556 cropped image dataset, weight mapping, augmentation, and artifact oversampling. Their model was trained on oversampled sub-images containing artifacts to improve model robustness against false positives. Second, they implemented a custom weight map that emphasizes the borders between touching cells, so the model could better separate clumped neuronal cells. We also retested Morelli and Clissa's c-ResUnet model with a reduced-data variant with no augmentation performed on the data and a smaller sample size of 500 samples to simulate limited training conditions due to compute limitations and real-world data limitations for many types of cells. This model did not use custom weight maps for the loss function or artifact oversampling, but instead used the weighted loss function described in Equation 1. We next introduced attention gates and experimented with both ELU and ReLU activations to assess their effect on segmentation performance, especially as the original Attention Unet paper used ReLU. We also increased the training sample size to 600 samples and used data augmentation here, to allow for a fair comparison to a non-attention based baseline model, the simplified c-Resunet, especially because attention-based models often require more data to successfully learn. Due to compute limitations here, we had to decrease the batch size to 4 instead of 8. The ELU-activated attention model was introduced to enhance feature flow and gradient stability, especially as the original c-ResUnet paper used this and ReLU can cause dead neurons in deep architectures. The ReLU variant however served as a good attention baseline for comparison. Augmentations, including shifts, flips, zooms, and rotations, were applied to help the model generalize better to low-sample scenarios in the attention models. The details of these models are shown in Table 1.

To refine the predicted binary masks and improve

cell segmentation, we applied a post-processing pipeline, inspired by Morelli and Clissa's [1]. First, we removed small holes and filtered out objects below a minimum size threshold of 40 to eliminate noise and artifacts. We then computed a distance transform on the cleaned mask to emphasize the centers of objects. Local maxima were detected in this distance map to serve as markers for individual instances. Using these markers, we applied watershed segmentation on the distance map to separate overlapping or touching regions, effectively delineating individual cells. This process produced cleaner masks and improved cell segmentation accuracy, especially in crowded regions.

To then use these masks to evaluate model performance, we could identify the center of each detected cell as the center of the bounding box around each clump of white pixels after post-processing. A predicted object was considered a true positive if its center lay within 40 pixels of a ground truth object center. This method used cell localization rather than strict pixel-wise overlap, which made it particularly suitable for biological cell segmentation tasks. This allowed us to compute true positives (TP), false positives (FP), and false negatives (FN).

Hence, throughout training and testing, we could evaluate our model's progress using commonly used metrics for segmentation: F1 score, Mean Absolute Error (MAE), Median Absolute Error (MedAE), Mean Percentage Error (MPE), Accuracy, Precision, and Recall.

The F1 score for binary masks, is defined as:

$$\text{F1} = \frac{2 \times \text{True Positive}}{2 \times \text{True Positive} + \text{False Positive} + \text{False Negative}} \tag{2}$$

It is often preferred in medical image segmentation because it balances Precision and Recall measurements. It can also be defined as:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Accuracy is the ratio of correctly predicted cells to the total number of cells:

$$\text{Accuracy} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive} + \text{False Negative}} \tag{4}$$

In this case, we ignore the background, also called true negative, as most of the pixels are background, and if considered accuracy will come out very high, even when the model doesn't actually predict all the positives correctly.

$$\text{Mean Absolute Error (MAE)} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{5}$$

where:

y_i = Ground truth value for sample i

$\hat{y}_i$ = Predicted value for sample i

N = Total number of samples

$$\text{Median Absolute Error (MedAE)} = \text{median}\left(|y_i - \hat{y}_i|\right) \tag{6}$$

$$\text{Mean Percentage Error (MPE)} = \frac{100\%}{N} \sum_{i=1}^{N} \frac{y_i - \hat{y}_i}{y_i} \tag{7}$$

Precision measures how many predicted positives are actually correct:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{8}$$

Recall measures the ratio of how many positives were actually detected:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{9}$$

The final output visualization of the model is a heat map of probabilities for each image, and after it is generated, we can use this heat map to visually evaluate failure modes and visually identify in which cases certain models perform better than others.

3. Experiments and Results

We evaluated the performance of the four variations of the c-ResUnet architecture through systematic experimentation. This section summarizes key observations from training behavior, hyperparameter tuning, threshold analysis, and post-processing.

3.1. Training Dynamics

The introduction of attention gates appeared to alter the convergence behavior of the model. The c-ResUnet with ELU and attention gating converged after 45 epochs, compared to 66 epochs for the baseline model without attention. When using ReLU with attention, convergence was even faster, stopping at 31 epochs. In each case, early stopping was triggered due to a plateau in validation loss, suggesting that the addition of attention accelerates learning but may also cause earlier overfitting if regularization or data diversity is limited. Training loss and validation loss remained close until the plateau for all 3 models we trained.

3.2. Learning Rate and Threshold Tuning

We initially experimented with several learning rates for 10 epochs each. However, $1e^{-4}$ consistently produced the best results in terms of F-1 scores, aligning with the prior model defined by Morelli and Clissa [1]. Our training

pipeline included learning rate reduction on plateau, which likely contributed to consistent performance across variations.

Threshold selection played a critical role in improving the performance of our models, as evaluated by the F1 score. The threshold is what was used to convert the probability maps outputted by the model to a binary segmentation map. As shown in Figure 3, we evaluated threshold values across a range to find the maximum F1 score, as we wanted to minimize the trade-off between precision and recall. The selected thresholds (0.6 for the official c-ResUnet model, 0.2 for the simple c-ResUnet, 0.3 for c-ResUnet ELU with attention, and 0.4 for c-ResUnet ReLU with attention) were chosen based on the peak F1 score on the validation set.

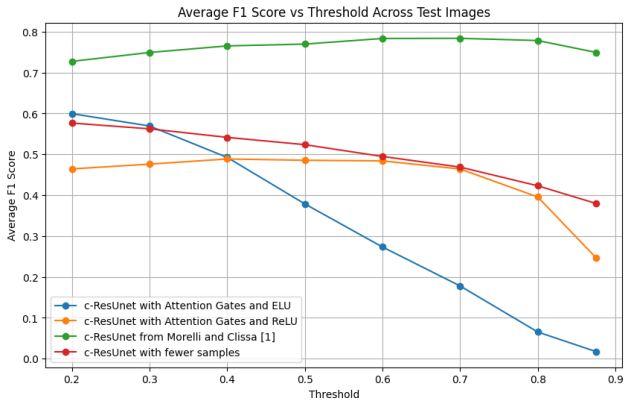

Figure 3. Hyperparameter tuning – searching for threshold values for each model

3.3. Distance Parameter Tuning in Post-processing

Our post-processing pipeline included center-based matching of predicted and ground truth cell segments, as described in Methods. The distance threshold for counting a true positive was tuned experimentally. A value of 40 pixels was selected after comparing multiple values, as it best balanced finding true positives with minimizing false positives due to common issues in a model like cell merging, for example.

3.4. Evaluating Performance Metrics

To assess the effectiveness of each model configuration, we computed key performance metrics as shown in the tables. These metrics provide a comprehensive view of both detection accuracy and counting precision.

The metrics for ELU-based c-ResUnet with attention gating model were very similar to our baseline simple c-ResUnet. They both have extremely simlar F1 scores of

Metric	Value
F1 Score	0.7835
MAE	2.7429
MedAE	2.0000
MPE	-0.0282
Accuracy	0.6441
Precision	0.8074
Recall	0.7610

Table 2. Performance metrics for the pre-generated model from [1]. A c-ResUNet trained with weight maps, artifact oversampling, and larger sample size of 2256 with additional augmentation. Threshold = 0.6

Metric	Value
F1 Score	0.5766
MAE	4.9143
MedAE	2.0000
MPE	-0.1074
Accuracy	0.4051
Precision	0.6821
Recall	0.4994

Table 3. Performance metrics for c-ResUNet with no weight maps or artifact oversampling. Threshold = 0.2, Sample size = 500, Batch size = 8

Metric	Value (ELU)	Value (ReLU)
F1 Score	0.5694	0.4886
MAE	5.6000	6.0429
MedAE	2.0000	4.0000
MPE	-0.3324	0.1263
Accuracy	0.3980	0.3233
Precision	0.7697	0.5428
Recall	0.4518	0.4443

Table 4. Performance metrics comparison of c-ResUNet with Attention Gates using ELU vs. ReLU activations.
Threshold = 0.2 for ELU and 0.4 for ReLU, Sample size = 600 with train-time data augmentation, Batch size = 4

0.5694 and 0.5766. However, notably the attention-based model with ELU has a slightly higher precision 0.7697 compared to 0.6821, which could make sense based on how attention might help the model learn how to ignore noisier parts of the image. This could also explain why there is a steep plummet in F1 scores with threshold values in Figure 3 – the model only segments cells its highly confident about. This less confident approach could also be why there is a slight difference in the MAE between the 2 models, with the attention-based model having an MAE of 5.6 cells compared to the other model's MAE of 4.9. Some of the limitations of the newly trained models may stem from dataset size and variability, as evidenced by the significant performance gap in every metric between our models and the pre-generated model from Morelli and Clissa [1], which benefited from a substantially larger dataset (2256

images), weight maps, and targeted oversampling of artifact-containing regions. These results highlight how both model architecture and data availability play critical roles in achieving robust performance.

Both the ELU and ReLU-based attention gating c-ResUnet models achieved similar F1 Scores (0.5694 for ELU vs. 0.4886 for ReLU). However, the ReLU based model did perform worse in every metric. The ReLU variant exhibited lower precision and higher MAE, suggesting it was more prone to over-segmentation, i.e. detecting cell regions where none existed, possibly due to ReLU's tendency to generate sparse but high-activation outputs. This behavior aligns with the observed positive mean percentage error (MPE = 0.1263), indicating a consistent overestimation in predicted cell counts.

3.5. Qualitative Image Analysis

In the qualitative results, as shown in figures 5 and 6, we observe that the attention-based model using ELU performs better in cluttered regions where cells are densely packed. Unlike the baseline model, which tends to overpredict and merge adjacent cells, the attention-based variant appears more conservative in its predictions. This restraint likely helps it distinguish between tightly grouped cells, resulting in improved segmentation in these challenging scenarios. This behavior suggests that attention gates help the model focus on cell boundaries and reduce over-segmentation.

In Figure 7, we observe a failure mode of the ReLU-based attention-gated c-ResUNet, where the model predicts too strongly around cell regions, causing neighboring cells to merge together.

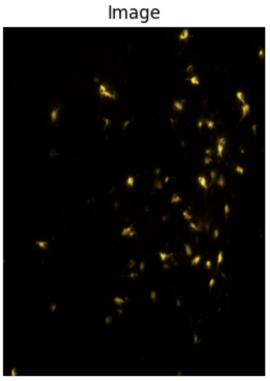

Figure 4. Image of cluttered cells

Figures 8–10 highlight another challenge: images with bright or noisy backgrounds. In Figure 9, the standard c-ResUNet performs slightly better by producing a stronger and more distinct probability map. However, as shown in

Figure 5. Cell Segmentation Visualization for cluttered cells image using simple c-ResUnet model

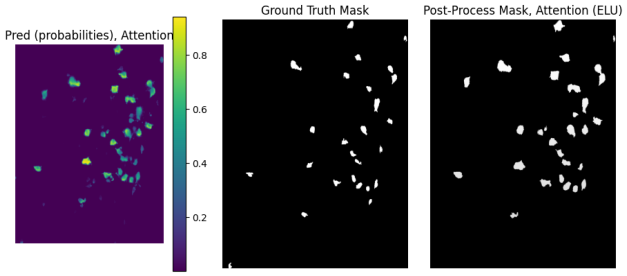

Figure 6. Cell Segmentation Visualization for cluttered cells image using c-ResUnet with ELU and attention gates. On the right side, it is clear the cells are separated more.

Figure 7. Cell Segmentation Visualization for cluttered cells image using c-ResUnet with ReLU and attention gates. Here, it is clear that the model shows high activation but tends to merge nearby cells, indicating difficulty in distinguishing densely packed regions.

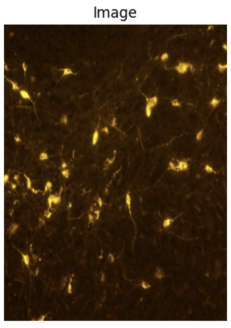

Figure 8. Image of cells with bright background

Figure 9. Cell Segmentation Visualization for bright/noisy cells image using simple c-ResUnet model. The probability map is a little stronger than the attention one below.

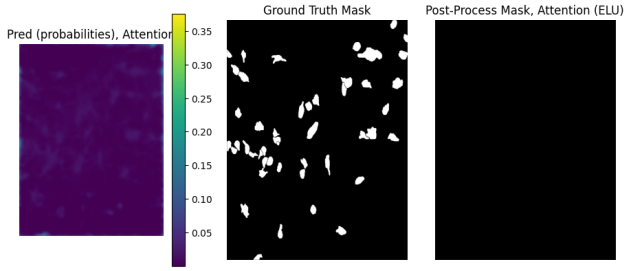

Figure 10. Cell Segmentation Visualization for bright/noisy cells image using c-ResUnet with ReLU and attention gates. The attention model fails here.

Figure 10, the attention-based model fails to accurately delineate cells under these conditions, suggesting that attention mechanisms may be more sensitive to image artifacts or brightness variations.

4. Conclusion

In this paper, we evaluated variations of the c-ResUNet architecture for cell segmentation, including the integration of attention gates and different activation functions. Our experiments demonstrate that while attention mechanisms can improve segmentation in cluttered cell environments by enhancing focus on salient boundaries, they are also more sensitive to data limitations and image artifacts. We observed that ELU activations provided slightly more stable performance compared to ReLU, which sometimes led to overconfident predictions and merged cell regions. Quantitative metrics such as F1 Score, MAE, and precision showed the tradeoffs between models, with the original paper's model still outperforming ours, highlighting the importance of large, well-augmented training sets.

4.1. Further Work

To further improve segmentation performance, future work could explore using Leaky ReLU as an activation function, which may help mitigate the vanishing gradient issue seen with standard ReLU, especially in cluttered regions. We can also try a different optimizer such as SGD to see if it improves performance. Additionally, training on the full sample dataset could allow the model to generalize better, particularly for attention-based architectures that rely on learning spatial context from lots of data. Lastly, integrating Vision Transformers (ViTs) [3], which have shown promise in biological image segmentation tasks due to their ability to model long-range dependencies, could provide a strong alternative to convolution-based architectures like c-ResUnet.

5. References

[1] Morelli, R., & Clissa, L. (2021, November 25). Automating cell counting in fluorescent microscopy through deep learning with c-ResUnet. *Scientific Reports*, *11*(1), 22920.

[2] Prangemeier, T., Reich, C., & Koeppl, H. (2020, November 19). Attention-Based Transformers for Instance Segmentation of Cells in Microstructures. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

[3] Santiago, J., Strohle, T., Rodriguez-Sanchez, A., & Breu, R. (2024, December 3). Vision Transformers for Weakly-Supervised Microorganism Enumeration.

[4] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

[5] Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T. (2018). Visualizing the loss landscape of neural nets. Advances in neural information processing systems.

[6] Morelli, R., & Clissa, L. (2021). *cell_counting_yellow* [Source code]. GitHub. Retrieved June 4, 2025, from `https://github.com/robomorelli/cell_counting_yellow`