CENTERSAM: FULLY AUTOMATIC PROMPT FOR DENSE NUCLEUS SEGMENTATION

Yiming Li[1], *Hanchi Ren*[1], *Jingjing Deng*[2], *Xiaoke Ma*[3], *Xianghua Xie*[1]

[1]Department of Computer Science, Swansea University, Swansea, United Kingdom
[2]Department of Computer Science, Durham University, Durham, United Kingdom
[3]Department of Computer Science, Xidian University, Xi'an, China

ABSTRACT

Nucleus segmentation is a vitally important task in biomedical image analysis which leads to multiple applications such as cellular behavior study, tumor detection and cancer diagnosis. However, challenges, such as ambiguous boundary for touching or overlapping nuclei often exist. This paper presents a dense nucleus segmentation method, namely CenterSAM combining the advantages from CenterNet and Segment Anything Model (SAM). It allows fully automatic prompting segmentation without prior knowledge enabling accurate and generalizable nucleus segmentation for biomedical images. Comprehensive evaluations of proposed method are performed on three nucleus segment benchmarks. The results highlight CenterSAM significantly out-performs the second best method by 5.3% on Dice Similarity Coefficient (DSC) in dense nucleus scenarios, meanwhile achieves competitive results on the sparse nucleus segmentation task. The code has been made publicly available[1].

Index Terms— Nucleus segmentation, CenterNet, Segment anything model.

1. INTRODUCTION

Nucleus segmentation plays an indispensable role in biomedical image analysis which contributes to numerous applications ranging from disease detection to drug discovery in biomedical research and clinical diagnostics, such as lesion determination and disease stage diagnosis. However, nucleus segmentation remains a challenge due to the complexity and variability of nucleus structures in size, shape, and appearance. Particularly, instance segmentation of nuclei on cell image requires delineating individual nuclei with precise boundaries, whereas the cells are often adhesive and boundaries are ambiguous due to the diversity in imaging modalities, low contrast inherent to tissue images, indistinct nucleus margins [1], and juxtaposition or overlapping of tissues [2]. In addition, the presences of noise and artifacts as results of imaging acquisition, impose further difficulties [3].

Transformer based architecture demonstrates the potential of a unified segmentation model across natural image and biomedical image domains, such as Segment Anything Model (SAM) that demonstrates great generalization capacity on both natural images [4] and medical images [5]. Furthermore, works such as MedSAM [6] and Univer-Seg [7] were proposed, which show the efficacy of SAM based approach in biomedical domains. However, these methods either target at the sparse segmentation for organs or rely heavily on manually provided prompts for high-performance segmentation. SAM on medical images shows that among the three prompt types bounding box as the input provides better segmentation performance in the vast majority of cases, especially in dense contexts such as cell/nucleus segmentation tasks. A simple, efficient, and accurate network for automatic prompt generation would be ideal to boost SAM-like model segmentation performance. CenterNet [8] has attracted public attention in the field that formulates the detection task as predictions of object centroid, aligning with the specific requirements of SAM and showed higher detection accuracy with faster training and inference. This anchor-free method is lightweight and versatile, allowing for adjustments based on varying situations and data volumes. Inspired by these observations, in this paper, we propose a two-stage method for dense nuclei segmentation, namely CenterSAM, that combines CenterNet as an automatic prompt generator and SAM as a precise segmenter. Our contributions are three-folds:

1. The proposed method eliminates the reliance on manual prompts, which are typically required by SAM-like models to achieve high-precision segmentation results.
2. Comprehensive experiments show the enhanced robustness and accuracy of CenterSAM on several nucleus segmentation benchmarks. It outperforms other competitors particularly in dense tissue scenario.
3. We further analyzed the size distribution and adhesion conditions of nuclei across multiple benchmark datasets, shedding a light on the rationale of performance gain in such densely packed scenarios for nucleus instance segmentation tasks.

2. PROPOSED METHOD

In this section, we present the technical details of the proposed CenterSAM model shown in Fig. 1, where Center-

[1] https://github.com/Rand2AI/CenterSAM.

Net based detector on the top produces box prompt for SAM based segmenter at the bottom.

2.1. Nucleis Detection

For an input image $I \in R^{W\times H\times 3}$, the detector first reduces the size of input by a factor of R = 4, and then pass it through two consecutive hourglass modules to produce the feature maps for detection heads. We choose stacked hourglass network [9] as the backbone for the detection stage, as it typically exhibits higher performance compared to common used detection networks such as Deep Layer Aggregation (DLA) [10]. Building on the concept of a bounding box design centered around the object's center point, three heads have been developed to predict the heatmap of object centroid, bounding box width-height, and bounding box offset respectively, where the losses for each head are defined as L_{H}, L_{wh} and L_{off} correspondingly.

For heatmap prediction, we represent a predicted keypoint by $\hat{Y}_{x,y}$ = 1 and $\hat{Y}_{x,y}$ = 0 as for background. Let $p \in R^2$ stand for the coordinate of a center keypoint of ground truth at the original image, we first project it to the coarse scale $\hat{p} = \lfloor \frac{p}{R} \rfloor$, then a Gaussian kernel $Y_{x,y} = \exp\left(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2}\right)$ is applied to this low-resolution ground truth heatmap, where σ_p is the size-adaptive standard deviation [11] of each instance. The loss of heatmap is defined as a combination of pixel-wise logistic regression with focal loss:

$$L_H = \frac{-1}{N}\sum_{x,y}\begin{cases}\left(1-\hat{Y}_{x,y}\right)^{\alpha}\log\left(\hat{Y}_{x,y}\right), & Y_{x,y}=1\\ \left(1-Y_{x,y}\right)^{\beta}\left(\hat{Y}_{x,y}\right)^{\alpha} & \\ \log\left(1-\hat{Y}_{x,y}\right), & \text{otherwise}\end{cases} \tag{1}$$

where N is the number of keypoints in an image and α and β are set to 2 and 4 following [8].

For bounding box prediction, let (x_1, y_1, x_2, y_2) be the bounding box of one instance J, the center point can be represented by $c_j = \left(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2}\right)$ and the size of the instance as $s_j = (x_2 - x_1, y_2 - y_1)$, for each prediction $\hat{S} \in \mathcal{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$, the L1 loss is calculated:

$$L_{\mathrm{wh}} = \frac{1}{N}\sum_{j=1}^{N}|\hat{S}_{c_j} - s_j| \tag{2}$$

In order to correct the offset error of each center point introduced by striding operation due to rounding down, a local offset $\hat{O} \in \mathcal{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$ is predicted and the smooth L1 loss is used for training. The loss function is defined as follow:

$$L_{\mathrm{off}} = \frac{1}{N}\sum_{p}\mathrm{SmoothL1}\left|\hat{O}_{\tilde{p}} - \left(\frac{p}{R} - \tilde{p}\right)\right| \tag{3}$$

Thus the overall loss function L_{det} for detection stage is defined in the following manner:

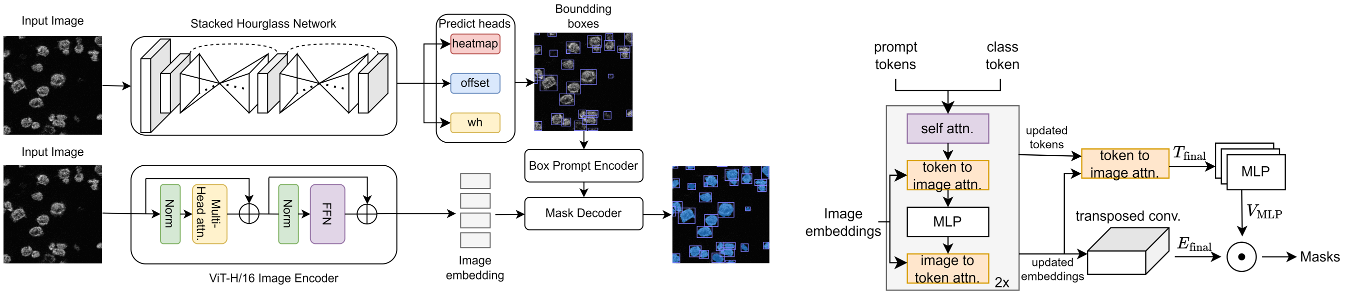
$$L_{\mathrm{det}} = L_{\mathrm{H}} + \lambda_{\mathrm{wh}} \cdot L_{\mathrm{wh}} + \lambda_{\mathrm{off}} \cdot L_{\mathrm{off}} \tag{4}$$

where λ_{wh} and λ_{off} are set to 0.1 and 1 if not otherwise specified. In the inference stage, the detection can be obtained given the predictions from three heads. First, the heatmap outputs are filtered to select points at a confidence score by a predefined threshold θ using non-maxima suppression (NMS). After the coordinate output of the predicted object center is determined, offsets are then added to the corresponding center coordinates followed by calculating the bounding box's region using the predicted width and height. Finally, the detected object bounding box can be represented with $(x_{\mathrm{min}}, y_{\mathrm{min}}, x_{\mathrm{max}}, y_{\mathrm{max}})$ that can be fed into the segmenter described in the following section.

2.2. Mask Segmentation

Image & Prompt Encoder: For the same pre-processed input image $I \in R^{W\times H\times 3}$, the segmenter resizes its spatial resolution to 1024×1024 and uses ViT-H/16 [12] as the image encoder. It produces a down-scaled (by a factor of 16) embedding with 14×14 windowed attention. To decrease the channel dimension, a 1×1 convolution is used to compress it down to 256 channels, followed by another convolution of filter size 3×3, also with the same number of channel output. Layer normalization is applied after each convolution. The image encoder was trained on SA-1B [4], where image embeddings that are calculated using the ViT-H/16 image encoder will be passed to the masks decoder. The bounding boxes are processed by the prompt encoder as follows: First, the original coordinates or input bounding box will be shift 0.5 pixel to ensure the coordinates point located at the center of the pixel; Then coordinates of "top-left corner" and "bottom-right corner" are passed to positional encoding module that utilizes random spatial frequencies to generate unique embedding to distinguish different positions in the image, which provides the model with rich spatial information; To further enhance the representation of the boxes prompt, an additional positional embedding weight is given to both "top-left corner" and "bottom-right corner", which indicates both the top left and bottom right corners of the box are specifically represented in the embedding space. This strategy further emphasizes the importance of these two key points that define the bounding box of target objects.

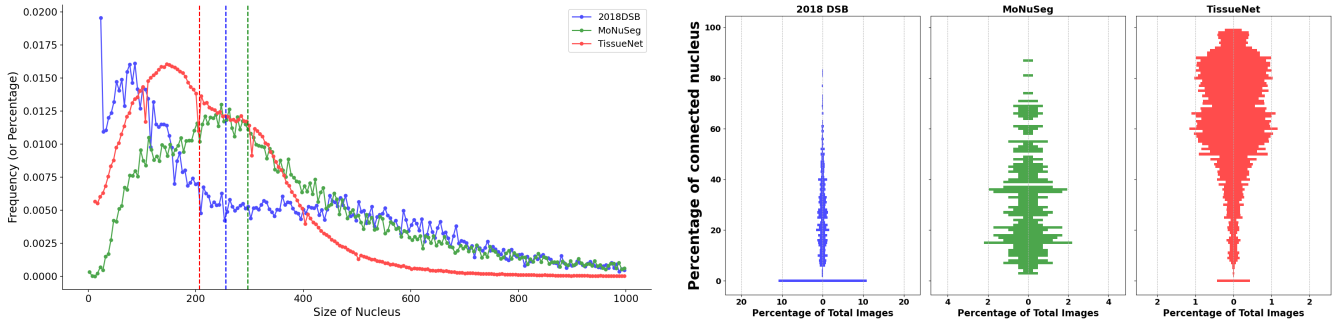
Mask Decoder: A learnable output token embedding is introduced which is similar to the [class] token in [12], where we refer this [class] token together with prompt tokens as "tokens". A single mask decoder layer (see Fig. 1(b)) first performs self-attention within tokens, followed by a cross-attention from tokens to the image embedding and versa vice. An element-wise Multi-Layer Perceptron (MLP) with residual connection is placed between two cross-attention. For each attention layer, we add the positional encodings to the image embedding and re-added the complete set of original prompt tokens along with their positional encodings to the updated tokens, to further tighten the location information and



(a) CenterSAM framework

(b) Mask decoder

Fig. 1: (a) The proposed CenterSAM framework for dense nucleus segmentation. CenterNet based detector on the top produces box prompts for SAM based segmenter at the bottom. (b) The architecture of mask decoder.



(a) Distribution of nucleus sizes.

(b) Distribution of connected nucleus.

Fig. 2: The statistics of nucleus cell. (a) the distribution of nucleus size and (b) the distribution of connected nucleus, where the dashed lines in (a) indicate the median values.

the corresponding prompt token. The updated tokens and embedding of the first decoder layer will then feed into another decoder layer with the same structure. After two decoder layers, the token embeddings will be updated once more by performing attention from output tokens to image embeddings, result marked as T_{final}. Meanwhile, the output image embeddings from two-layer decoder is upsampled by a factor of 4 using two transposed convolutional layers. The size of convolution kernel is set to 2×2, with a stride of 2 followed by a Gaussian Error Linear Unit (GELU) activation. The final image embedding is referred as E_{final} with channel dimension of $D = 2048$. The T_{final} is fed into a small 3-layer MLP to produce a vector V_{MLP} with same channel dimension D. The mask prediction is obtained by performing a spatial point-wise multiplication between the E_{final} and V_{MLP}.

3. EVALUATION AND DISCUSSION

3.1. Dataset and Implementation

In order to evaluate the performance of the proposed CenterSAM comprehensively, we used three standard benchmarks listed in Table 1, including MoNuSeg [22] as a representative for small dataset, the 2018 Data Science Bowl dataset (2018 DSB) [23] for medium-sized dataset, and TissueNet [21] for large dataset. These datasets cover different organs (such as Breast, Kidney, Liver, Prostate, Bladder, Colon and Stomach), various species (including humans, mice, and macaques), and imaging techniques (such as bright-field and fluorescence).

Table 1: The benchmark datasets.

Dataset	Imaging	Images	Resolution	Annotation
MoNuSeg	H&E	51	1000*1000	32,217
2018 DSB	multimodal	670	Variable	29,461
TissueNet	multiplexed	6990	512*512	~1.2M

CenterSAM was trained on each dataset individually using a GeForce RTX 3090 graphics card. For MoNuSeg and TissueNet dataset, we followed the default split of the dataset. For 2018 Data Science Bowl dataset, we randomly split 670 labeled images into an 80-10-10 train-val-test proportion. We followed the process from Cellpose [24] and cropped the original 1000×1000 image from MoNuSeg into 9 non-overlap images. Contrast Limited Adaptive Histogram Equalization (CLAHE)[25] was applied to all images with a threshold of 2 and adaptive grid size of 8×8, then the images are all resized to 512×512 before feeding into the models. We increased the default probabilities of multiple augmentations of CenterNet to enhance the diversity of training images. To be more specific, the probability for applying shift, scale, rotate and flip augmentation for an input image is increased to 0.5. We set the learning rate at 1.25×10^{-4} with batch

Table 2: Quantitative results of comparison against State-Of-The-Art (SOTA) methods. The best results are highlighted in bold.

2018 Data Science Bowl			**MoNuSeg**			**TissueNet**		
Method	DSC(%)↑	mIoU(%)↑	Method	DSC(%)↑	AJI(%)↑	Method	DSC(%)↑	SEG(%)↑
UNet++	91.10	83.70	UNet	74.56	60.22	Detectron2	75.50	78.00
Deeplabv3+ [13]	88.80	83.70	UNet++ [16]	80.33	67.30	Cellulus [19]	64.10	52.40
SSFormer-S [14]	92.50	86.50	MAE [17]	73.68	58.62	StarDist [20]	59.40	38.20
DuAT [15]	**92.60**	**87.00**	MDM [18]	81.01	68.25	Mesmer [21]	83.40	77.20
CenterSAM	92.20	86.60	**CenterSAM**	**81.95**	**68.75**	**CenterSAM**	**88.70**	**79.50**

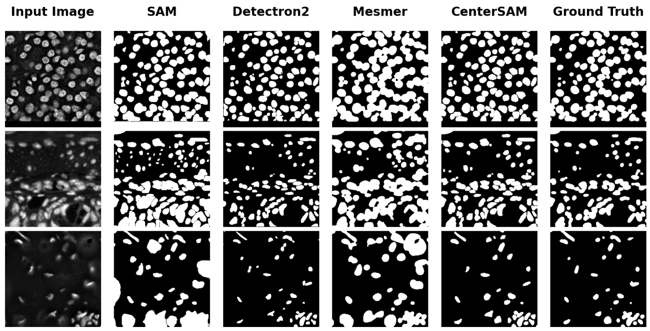


Fig. 3: Qualitative results of comparison against selected methods on different datasets.

Table 3: The density of nucleus cell.

Dataset	Total	Connected	Proportion	Mean
2018DSB	29,461	8,566	29.08%	22.69%
MoNuSeg	32,217	12,050	37.40%	32.61%
TissueNet	1,286,856	944,651	73.41%	64.37%

size of 32 and train epochs of 140. A learning rate decay was scheduled by a scale of 10 at epochs 90 and 120. The detector model was trained from scratch, while a pre-trained segmenter is used. For MoNuSeg, ResNet-18 is used instead of the default stacked hourglass network to mitigate the risk of over-fitting due to small training sample. The max number of output objects K was increased from 100 to 1000 to avoid missing small instance, and θ=0.1 for NMS threshold is used. To evaluate performance differences, we adopt Dice Score (DSC), mean IoU (mIoU), Aggregated Jaccard Index (AJI) and SEG score [26] as the evaluation metrics, where predictions with an IoU equal or greater than 0.5 will be considered as a successful match for DSC measurement.

3.2. Experimental Result

Table 2 and Fig. 3 show the comparison results against SOTA methods on different datasets. The proposed CenterSAM achieves the highest accuracy on both MoNuSeg and TissueNet datasets. Particularly, our method shows significant improvement over Mesmer (the second best) on TissueNet in terms of both DSC and SEG scores by 5.3% and 2.3% respectively. It is worth noting both Mesmer and StarDist require shape prior knowledge which is no-need for our model. On the 2018 DSB dataset, CenterSAM is highly competitive with only a 0.4% gap in DSC and AJI scores compared to DuAT.

We measured the inference speed on the TissueNet dataset. Detectron2 is the fastest with $3.96 \times 10^{-2} \pm 3.10 \times 10^{-3}$s per image, followed by StarDist with $6.21 \times 10^{-2} \pm 2.37 \times 10^{-2}$ CenterSAM with $7.05 \times 10^{-2} \pm 2.16 \times 10^{-3}$ and Mesmer with $1.02 \times 10^{-1} \pm 8.09 \times 10^{-3}$. CenterSAM demonstrates high efficiency with the same accuracy while having the lowest standard deviation.

To delve into the performance of CenterSAM across three datasets, we further analyzed the cell nucleus sizes and density of each datasets (see Fig. 2) to further explore the optimal scenario for applying CenterSAM. 2018 DSB contains a significantly higher proportion of extremely small nuclei but have a higher median number than TissueNet dataset. Our approach is ranked the third place with merely 0.3% lower but 0.1% higher than the second place in terms of DSC and mIoU scores respectively. It suggests that CenterSAM is capable of capturing extremely tiny biomedical instances, while the segmenter was trained on SA-1B dataset that contains mainly natural images. We further calculated the average number of nucleis per image and the percentage of nucleis that connected with another nucleis from randomly cropped the image with a fixed size of 512×512. The visualization of the results is shown in Fig. 2(b). The 3 shows the quantitative statistics of nucleus density. TissueNet has 73.41% of nuclei exhibiting connection or overlap with at least one neighboring nucleus, whereas CenterSAM outperforms the SOTA by a significant margin (+5.3% in DSC and +2.3% in SEG). Such superiority in performance is also evident in the MonuSeg dataset that has an approximate 37.40% rate of nuclei connection. Our approach surpasses the current SOTA to a notable extent. On the 2018 DSB dataset, where the rate of nuclei contact drops to 29.08%, our method is slightly behind the SOTA.

4. CONCLUSION

In this paper, we proposed CenterSAM, a fully automatic nucleus instance segmentation method with prompt based model. It requires no shape or appearance prior knowledge and manual prompt. The results on multiple benchmarks underscore its robustness in scenarios where dense nucleus structures and high overlap rates are presented. Such scenarios are inherently challenging due to the ambiguity in boundary demarcation, making nucleus segmentation a non-trivial task. The proposed method is more effective than SOTA suggesting its potential advantages in practical applications.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human and animal subject data that are publicly available in open access. Ethical approval was not required.

6. ACKNOWLEDGMENTS

All authors declare that they have no conflicts of interest. Xiaoke Ma is funded by the Key Research and Development Program of Shaanxi (Program No. 2021ZDLGY02-02).

7. REFERENCES

[1] Eli Gibson and et al., "Niftynet: a deep-learning platform for medical imaging," *Computer methods and programs in biomedicine*, vol. 158, pp. 113–122, 2018.

[2] Shujian Deng and et al., "Deep learning in digital pathology image analysis: a survey," *Frontiers of medicine*, vol. 14, pp. 470–487, 2020.

[3] Tomohiro Hayakawa and et al., "Computational nuclei segmentation methods in digital pathology: a survey," *Archives of Computational Methods in Engineering*, vol. 28, pp. 1–13, 2021.

[4] Alexander Kirillov and et al., "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[5] Guochen Ning and et al., "The potential of'segment anything'(sam) for universal intelligent ultrasound image guidance," *BioScience Trends*, 2023.

[6] Jun Ma and Bo Wang, "Segment anything in medical images," *arXiv preprint arXiv:2304.12306*, 2023.

[7] Victor Ion Butoi and et al., "Universeg: Universal medical image segmentation," *arXiv preprint arXiv:2304.06131*, 2023.

[8] Xingyi Zhou and et al., "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[9] Alejandro Newell and et al., "Stacked hourglass networks for human pose estimation," in *ECCV*. Springer, 2016, pp. 483–499.

[10] Fisher Yu and et al., "Deep layer aggregation," in *CVPR*, 2018, pp. 2403–2412.

[11] Hei Law and Jia Deng, "Cornernet: Detecting objects as paired keypoints," in *ECCV*, 2018, pp. 734–750.

[12] Alexey Dosovitskiy and et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] Liang-Chieh Chen and et al., "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV*, 2018, pp. 801–818.

[14] Jinfeng Wang and et al., "Stepwise feature fusion: Local guides global," in *MICCAI*. Springer, 2022, pp. 110–120.

[15] Feilong Tang and et al., "Duat: Dual-aggregation transformer network for medical image segmentation," *arXiv preprint arXiv:2212.11677*, 2022.

[16] Zongwei Zhou and et al., "Unet++: A nested u-net architecture for medical image segmentation," in *MICCAI*. Springer, 2018, pp. 3–11.

[17] Kaiming et al. He, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16000–16009.

[18] Zixuan Pan and et al., "Masked diffusion as self-supervised representation learner," *arXiv preprint arXiv:2308.05695*, 2023.

[19] Steffen Wolf and et al., "Unsupervised learning of object-centric embeddings for cell instance segmentation in microscopy images," in *ICCV*, 2023, pp. 21263–21272.

[20] Uwe Schmidt and et al., "Cell detection with star-convex polygons," in *MICCAI*. Springer, 2018, pp. 265–273.

[21] Noah F Greenwald and et al., "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning," *Nature biotechnology*, vol. 40, no. 4, pp. 555–565, 2022.

[22] Neeraj Kumar and et al., "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[23] Juan C Caicedo and et al., "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[24] Carsen Stringer and et al., "Cellpose: a generalist algorithm for cellular segmentation," *Nature methods*, vol. 18, no. 1, pp. 100–106, 2021.

[25] Ali M Reza, "Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 38, pp. 35–44, 2004.

[26] Vladimír Ulman and et al., "An objective comparison of cell-tracking algorithms," *Nature methods*, vol. 14, no. 12, pp. 1141–1152, 2017.