Multiscale Feature Learning and Cross Spatial Attention in Mask R-CNN for enhanced Cell Instance Segmentation

Dipankar Jiwani[1], Rachita Saha[1], Isha Sehrawat[1], Anubha Gupta[1,*], Anish Jain[1], Ritu Gupta[2]

[1]SBILab, Department of ECE, IIIT-Delhi, India
[2]Dr. BRA.IRCH, Laboratory Oncology Unit, AIIMS, New Delhi, India
Email: {dipankar19037, rachita19082, isha19046, anubha, anish22077}@iiitd.ac.in (D.J., R.S., I.S., A.G., A.J.); drritugupta@gmail.com (R.G.)
*Corresponding author

Manuscript received February 29, 2024; Revised June 9, 2024; accepted June 11, 2024; published October 16, 2024

***Abstract*—Cell instance segmentation in medical imaging is pivotal for advancing diagnostics and treatment. Despite the acknowledged importance of Mask R-CNN for this task, we observed challenges in effectively distinguishing some boundary pixels, particularly in scenarios where cells are in close proximity. To address these issues, this research introduces three key enhancements: 1) Multiscale Feature Learning (MSFL), 2) the Cross Spatial Attention Module (CSAM), and 3) Novel training of UNet for guiding the training of Mask RCNN through CSAM module. MSFL utilizes various features produced by the FPN backbone across different scales, thereby minimizing data loss and enhancing the overall representation of the region of inter- est. The lightweight CSAM significantly enhances segmentation results by harnessing the inherent segmentation capabilities of U-Net in the medical domain. This novel approach not only rectifies boundary errors, but also enhances accuracy and robustness in medical image analysis. Importantly, the adaptable CSAM seamlessly integrates into various models, increasing segmentation accuracy without a substantial impact on the model size. The efficacy of this approach is demonstrated through its application on two distinct cell segmentation datasets. Results demonstrate a notable increase of 2.66% mean Intersection over Union (mIOU) from the baseline on the SegPC dataset and a significant improvement by 1.86% in the mAP@[0.5:0.95] on the Yeast Cell dataset.**

***Keywords*—medical imaging, cell instance segmentation, Mask R-CNN, U-Net, spatial attention**

I. Introduction

Instance Segmentation is crucial in medicine, guiding treatment decisions by precisely locating, segmenting, and classifying abnormalities like cancer, tumors, or lesions [1]. This task involves classifying regions of interest, such as tumors, cells, organs, and tissues, and distinguishing between their instances. Medical datasets often present challenges like varied shapes of cells/structures, poor illumination, uneven staining, and complexities such as cell division and overlapping cells. Overcoming these challenges requires a resilient and robust representation capable of handling the intricacies within the medical domain. Cell segmentation involves diverse methodologies, from classical computer vision techniques to con- temporary deep learning paradigms. Some of the famous architectures for medical image segmentation have been the U-Net architecture [2] and its variants [3, 4]. UNet is a convolutional neural network (CNN) based encoder-decoder type architecture that has been successfully used in cell segmentation. The architecture, unlike the fully convolutional network (FCN), involves the concatenation of feature maps from the encoder layers to decoder layers, which helps in context information preservation and generation of smoother segmentation masks.

In the recent past, deep learning methods, exemplified by DeepMask [5] and SharpMask [6], have significantly enhanced capturing of spatial relationships. *Segment Anything* [7] is recognized for effectiveness despite constraints in segmenting fine structures, while Mask R-CNN [8] that has proven crucial in cell segmentation, also encounters challenges at boundaries. Models like Cascade R-CNN [9] and Hybrid Task Cascade [10], built upon Mask R-CNN, show promising results, but face challenges of training a higher number of parameters. This paper adopts Mask R-CNN for instance segmentation and proposes enhancements to address its limitations.

Attention mechanisms, such as visual attention [11], residual attention networks [12], SENet [13], and TransAttUnet [14], have gained prominence in image segmentation due to their efficacy in modeling the importance of features in a given task. These models incorporate attention elements, including spatial attention, channel attention, and relation functions, to enhance feature representation and leverage con- textual information. Recent developments such as BAM [15], CBAM [16], and DAN [17] fuse spatial and channel attention, while Efficient Channel Attention (ECA) [18] employs a local cross-channel interaction strategy without dimensionality reduction. ECA is noteworthy for its efficiency and significant performance improvements with only a limited number of parameters and computations [18]. This paper leverages the ECA module to enhance the Mask Head in Mask R-CNN [8]. Existing attention mechanisms rely on their inherent features, and thus fall short in capturing domain-specific intricacies, especially in medical imaging. One promising approach for enhancing these features involves strategically integrating domain-specialized models known for their excellence in specific contexts. Thus, to address the challenges at boundaries in cell segmentation, this work introduces a modification to the Mask R-CNN by introducing three novel ideas:

1) **Multiscale Feature Learning:** The FPN backbone in Mask R-CNN generates multiple features at different scales for each region proposal. However, traditional methods use only one feature based on scale matching. MSFL leverages all generated features simultaneously, and thus enhances learning, reduces potential data loss, and captures complementary information at various scales, fostering a more comprehensive representation

of the considered region.

2) **Cross-Spatial Attention Module:** In our proposed attention module, we enrich feature representation by collaboratively incorporating domain-specific promising models to extract more contextually relevant information. Notably, we leverage the U-Net architecture [19], known for its state-of-the-art performance in medical image segmentation tasks. This synergy harnesses the strengths of both models, promising improved performance and adapt- ability across diverse tasks and domains. Importantly, this approach is inherently generalizable across diverse domains, allowing integration of well-performing models from different domains to capitalize on their expertise. By adapting attention modulation based on guided attention maps from high-performing models, our method offers a flexible and effective means of incorporating prior knowledge, enhancing overall performance and adaptability.

3) **Novel training of U-Net Architecture**: We leveraged U-Net's strong performance in medical semantic segmentation to tackle instance segmentation. We performed fine- tuning of the U-Net model using our carefully curated dataset. The input images were designed to simulate ROI proposals on cell images, incorporating neighboring regions and overlapping cells. The ground truth annotations were structured to enable the model to distinguish between the primary cell within each proposed region and neighboring cells. This strategy effectively enhances the feature maps of U-Net, thereby assisting Mask RCNN in refining instance segmentation performance via the CSAM module.

II. Proposed Model

In the subsequent sections, we elucidate how we enhanced the baseline Mask-RCNN by introducing Multiscale Features Learning, the Cross Spatial Attention Module, and novel U-Net training.

A. Multiscale Feature Learning (MSFL)

Mask R-CNN is a two-stage detector network, initially generating region proposals and subsequently predicting class, bounding box, and mask for each proposal. Employing a Feature Pyramid Network-equipped backbone, it extracts image features across various scales. RPN is a Region proposal network that identifies proposals of regions of interest, with the Multiscale ROI Align providing a feature resized to predetermined dimensions. Despite having access to three-scale feature maps from the FPN backbone, the model opts for a single feature map based on scale matching potentially resulting in information loss. To improve the comprehension of visual content and to obtain more contextual information about the ROI, our solution involves utilizing all these multiscale feature maps. This issue is addressed through a novel mask head for Mask R-CNN, optimizing pixel-level mask prediction by incorporating all feature maps and enabling the model to autonomously determine their relative importance.

Our proposed mask head requires feature maps at various scales for each ROI. MSFL provides fixed-sized feature maps for ROIs at various scales. Initially, image feature maps sized 64×64, 128×128, and 256×256 from the Feature Pyramid Network (FPN) backbone are passed into the Multiscale ROI Align along with the ROI proposals. This process generates feature maps of fixed sizes 16×16, 32×32, and 64×64 corresponding to each input image feature map for the ROIs. These feature maps then go through separate convolution layers within the mask head. The resulting outputs are concatenated, combining the information from multiple scales, e.g., the 64×64-sized feature is pooled, and the 16×16-sized feature is transposed convoluted to achieve a uniform size of 32×32 (Fig. 1). This concatenated set of features is passed through the ECA network that learns weights to assign relative importance to feature maps of different scales extracted originally. Next, the set of feature maps undergoes two additional convolution layers. Following this, cross-spatial attention from U-Net is introduced to further enhance these features through the Cross Spatial Attention Module, which is elaborated upon in the subsequent discussion.

B. Cross Spatial Attention Module (CSAM)

The Cross Spatial Attention Module takes as input two crucial features: the Guiding feature and the Guided feature. The guiding feature is the segmentation map derived from a model recognized for its excellence in image segmentation within the specified domain. The guided feature corresponds to the output obtained after employing the ECA step in the MSFL process. Both features undergo separate Max and Average Pooling (Adaptive), individually adjusted to the size of the guided feature (32×32). Adaptive Pooling allows us to specify the output size, with the stride and kernel size automatically adjusted to suit the requirements. This adaptive approach ensures consistent output sizes regardless of input dimensions, which are then concatenated and fed through a convolution layer producing four channels as output. This process is repeated for both the guided and guiding features. Subsequently, the outputs of both the guided and guiding features are concatenated and further processed through a final convolution layer, followed by a sigmoid activation to yield the guided spatial feature attention map.

This spatial attention map is a synthesis of information from both the guided and guiding features, creating a comprehensive attention map for the guided feature. To ensure compatibility, the spatial attention map is extended (replicated along the channels) to match the channels of the guided feature. The two are then subjected to element-wise multiplication and returned. The architecture of the Cross Spatial Attention Module is depicted in Fig. 2.

C. Unified Model Architecture

We apply Multiscale Feature Learning (MSFL) to generate the guided feature, dynamically integrating features of diverse scales adjusted according to their individual significance. Additionally, we incorporate spatial attention from the U-Net model, known for its proficiency in analyzing medical images. A dedicated U-Net model is trained to generate segmentation masks for each ROI extracted from the dataset. The proposal region is cropped from the original image, and the U-Net model produces a segmentation mask of dimensions 32×32. This segmentation mask, post-convolution, serves as the guiding feature for the CSAM

module. It collaborates with our initial feature from the primary model which served as the guided feature. The output of the CSAM is integrated into our original guided feature, resulting in a significant enhancement, and after passing through two additional convolution layers, the final ROI mask is accurately predicted. The complete architecture of our unified model is illustrated in Fig. 1.

III. Experimental Setup

The primary dataset utilized in this study is the SegPC-2021 dataset [20–22], featuring microscopic images of bone marrow aspirate slides from patients diagnosed with Multiple Myeloma (MM). Captured in raw BMP format using two cameras affixed to the microscope, the images come in two sizes: 2040×1536 pixels and 2560×1920 pixels. These images have been stain normalized by the authors of the dataset [23, 24]. The dataset comprises a total of 409 images, with 120 allocated to the training set, 12 for validation, and 277 for the final test dataset. The dataset's objective is to facilitate the segmentation of each cell instance (nucleus + cytoplasm), labeled as Background: '0' and Cell: '255'. Training encompasses whole images, and augmentations, with the evaluation conducted on the entire microscopic images.

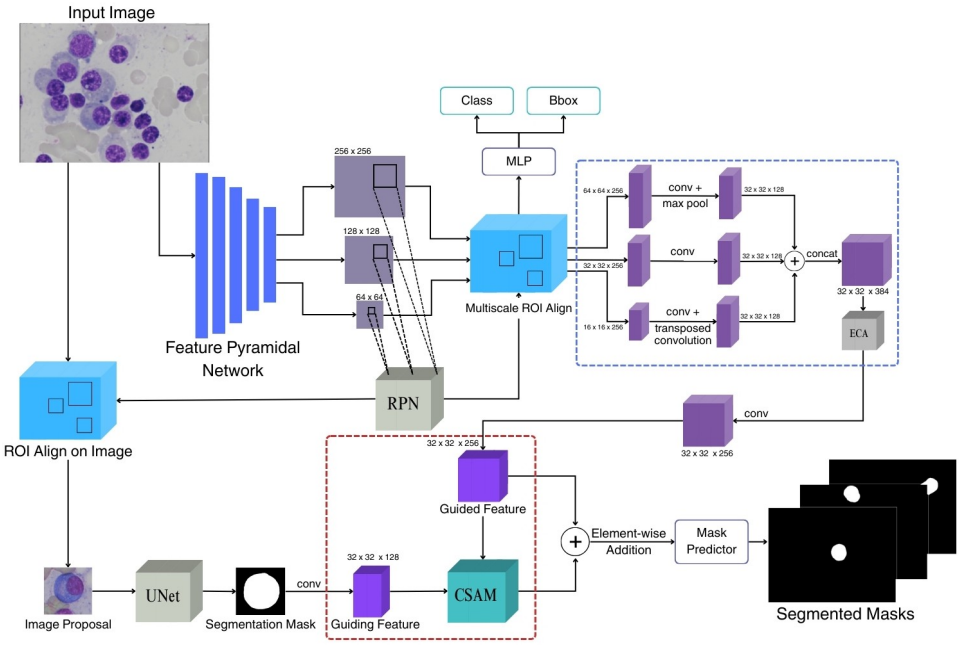


Fig. 1. **Proposed Model Architecture:** Our model employs Multiscale Feature Learning (highlighted in blue dotted lines) on ROIs proposed by the Multiscale ROI align, dynamically adjusting features of varying scales for improved segmentation accuracy. The Cross Spatial Attention Module (highlighted in red dotted lines) leverages U-Net's segmentation output, specifically trained for precise cell mask segmentation in response to ROIs. The updated feature facilitates efficient instance- wise segmented masks, embodying a comprehensive approach to instance segmentation in our unified model. The ECA module has been taken from the reference [18]. All the convolutional layers incorporate ReLU as the activation function.

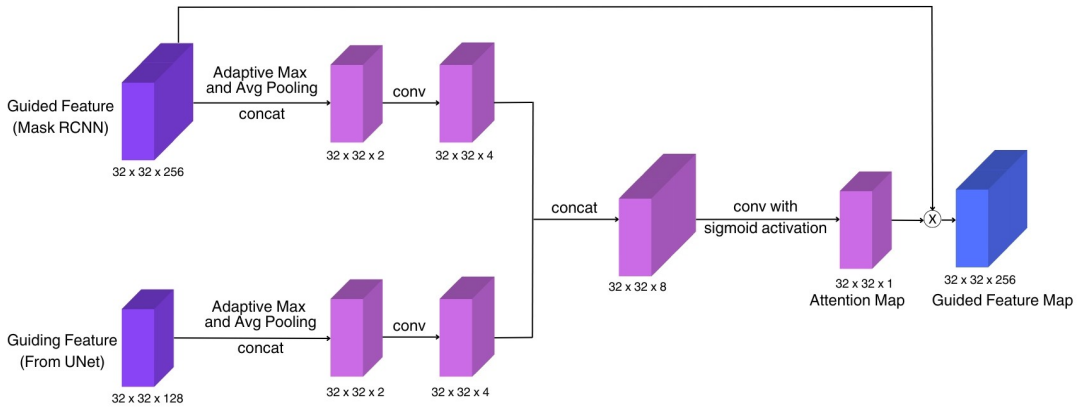


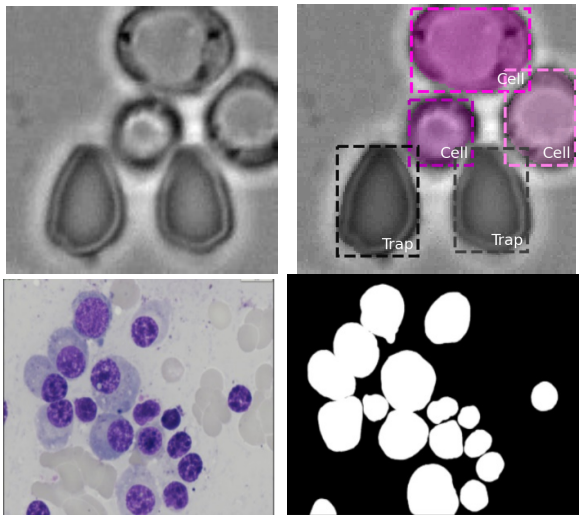
Fig. 2. **Cross Spatial Attention Module:** The CSAM improves instance segmentation by guiding the model with spatial attention, utilizing the segmentation output from a model recognized for its performance in the specific domain.

To assess the proposed model's generalizability, a second dataset of yeast cell segmentation in microstructures is utilized [25]. It comprises of 493 dense annotated microscopy images. This dataset provides pixel-wise instance segmentation labels for yeast cells and trap microstructures. Used as a validation benchmark, this second dataset enables an evaluation of our model's consistency and generalizability across diverse bio-logical scenarios. Some sample images of both the dataset are shown in Fig. 3.

A. Loss Function and Evaluation Metrics

The training process incorporates a composite loss, which encompasses classification, bounding box regression, and mask loss, following the definitions provided in [8]. To evaluate the segmentation algorithm's performance, we employ mIoU as the primary metric. IoU measures the overlap between ground truth and predicted segments. Given

the two classes (background and cell), IoU is calculated separately for each class. The mean IoU is then determined by averaging the IoUs for both classes, offering a comprehensive measure of segmentation accuracy.



(a) (b)

Fig. 3. Sample Images from the SegPC-2021 [20] and Yeast Cell Dataset [25]: (a) Input images (b) Corresponding Ground truth instance segmentation masks.

B. Training Details

1) MaskRCNN Training: During training, in order to make the input images of a uniform size, symmetrical zero padding was applied to make them of the dimension 2560×1920. Several data augmentation techniques were employed such as random horizontal and vertical flip with probability 0.3 and varying the hue, saturation, brightness and intensity values of images. We experimented with widely used backbone networks ResNet-50, ResNet-152, ResNext-101 and WideResNet-101 as well.

We employed an SGD optimizer with a momentum of 0.9 and a weight decay of 5e-4 during training. Additionally, we utilized a Cyclic Learning Rate Scheduler in triangular2 mode with a base learning rate of 1e5, a maximum learning rate of 0.006, and a step size of 5, while training on a batch size of 4. All models were implemented using the PyTorch framework and trained on Nvidia DGX GPUs.

2) U-Net Training: The U-Net model, employed for generating semantic segmentation masks that would serve as the guiding feature for the CSAM module, underwent dedicated training tailored for this specific task. To curate the dataset for ROI mask generation, an iterative process was initiated across all instance segmentation masks for each image in the SegPC dataset. This involved identifying the top, bottom, left, and right boundary pixels of the masks corresponding to each cell instance, with an additional value added to encompass neighboring cell regions. This modified region was cropped to serve as the ground truth for training. Concurrently, the corresponding region with identical boundary pixels was extracted from the training image. These cropped regions simulate ROI proposals similar to those generated by Mask R-CNN, encompassing not only the primary cell but also neighboring or overlapping cells.

However, our dataset design ensures that the U-Net is exclusively trained to disregard these additional cells and concentrate on predicting the primary cell within the ROI. Fig. 4 depicts training of U-Net. This meticulous approach elevates the model's proficiency in discerning and outlining the primary cell of interest, thereby refining precision in semantic segmentation within the designated ROI. This enhanced segmentation output serves as a robust guiding feature for the CSAM.

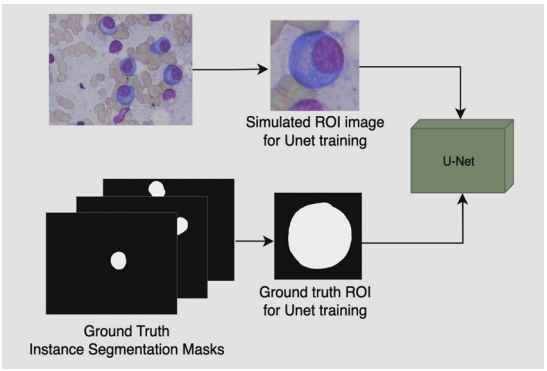


Fig. 4. **U-Net Training**: Simulated ROI proposals encompassing the primary cell as well as neighboring or overlapping cells are passed as input to fine-tune U-Net. The dataset ensures that U-Net is exclusively trained to disregard neighboring cells and predict the primary cell within the ROI.

IV. Results and Analysis

Within this section, we conduct a comprehensive ablation study, evaluating our model's performance on the SegPC-2021 challenge test set through mIoU metric. Our investigation delves into the influence of various factors, including the model backbone, training augmentations, MSFL, and the CSAM module. The results obtained are presented and summarized in Table 1.

Table 1. Experimental Results on SegPC dataset [20]: Comparison of baseline Mask R-CNN with the outcomes achieved with our proposed multiscale feature learning and cross-attention module

Model	Backbone	mIOU	Number of parameters
Mask RCNN	ResNet50	0.8712	43.9M
Mask RCNN+Augmentation	ResNet50	0.8841	43.9M
Mask RCNN	ResNeXt101	0.8900	107M
Mask RCNN+Augmentation	ResNeXt101	0.8925	107M
Mask RCNN+MSFL	ResNet50	0.9027	44.4M
Mask RCNN+MSFL+CSAM	ResNet50	**0.9107**	44.5M
Mask RCNN+MSFL	ResNeXt101	0.9072	107.5M
Mask RCNN+MSFL+CSAM	ResNeXt101	**0.9133**	107.6M

A. Ablation Study

1) Optimization Strategies and Backbone Selection: We use three key optimization strategies for refining the instance segmentation model. First, we implement various

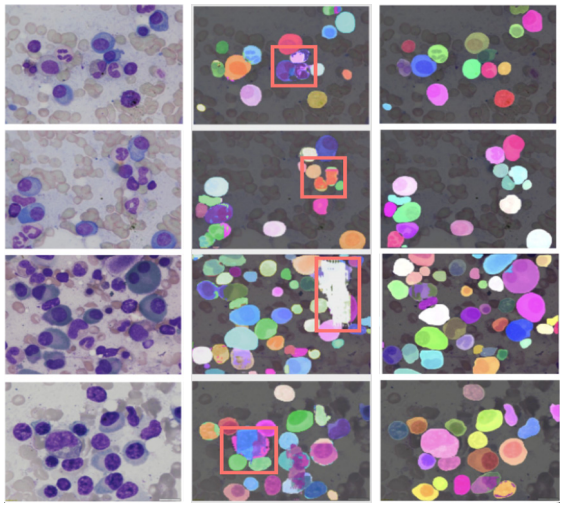
augmentation techniques, including changes in hue, saturation, brightness, intensity values, and random flips. Keeping other hyperparameters constant, these augmentations elevate the mIOU to 0.8841, showcasing a 1.3% mIOU improvement over the vanilla Mask RCNN. Considering the critical role of the backbone in Mask RCNN's performance, we experiment with ResNet50, ResNet152, ResNeXt101, and WideResNet101 backbones. We find that the ResNeXt101 backbone yielded the highest mIOU of 0.8900, outperforming its counterparts, while ResNet52 backbone performed second best. Hence, we have shown results with ResNeXt101 and ResNet52 backbones. Mask RCNN, equipped with the ResNeXt101 backbone, and dataset augmentations, attains an impressive mIOU of 0.8925. The rationale behind this improvement lies in the model's exposure to a more extensive range of variable and diverse samples, fostering increased robustness

2) Multiscale Feature Learning: The application of our multi-scale feature learning technique to the Mask RCNN model with the ResNet50 backbone significantly increases the test mIOU from 0.8841 to 0.9027. This noteworthy improvement highlights the potential loss of vital information when only one feature is selected from the FPN and others are discarded. The discarded features may contain crucial details contributing to enhanced segmentation. Consequently, our approach, involving attention-driven concatenation of all features, proves more effective, as substantiated by the achieved results.

3) Cross Spatial Attention Module: Mask RCNN, coupled with Multi-Scale Feature Learning alone, attains a respectable 0.9027 mIOU. However, upon incorporating the cross-spatial attention approach, we observe a significant boost in performance, achieving an mIOU of 0.9107 with ResNet50 backbone and 0.9133 with ResNeXt101 backbone. Leveraging the strengths of U-Net in medical-domain semantic segmentation, we produce guiding features for each region proposal which play a crucial role in directing the model on where to focus within the guided features generated for that specific region proposal from the FPN. Thus, employing the Cross Spatial Attention module enables us to harness the outstanding abilities of one model to guide another effectively.

B. Visualization of Test Outputs

We conduct a comparative analysis between the instance segmentation masks generated by Model A, based on Vanilla Mask RCNN with data augmentations and a ResNeXt101 backbone, and our proposed Model B with MSFL and CSAM. The results obtained on the test images are presented in Fig. 5. Our analysis reveals that Model A struggles to effectively distinguish between individual cell instances when they appear cluttered together, often predicting multiple overlapping masks. Consequently, the boundaries of the cells exhibit irregularities and lack coherence. In contrast, our proposed model produces smooth boundaries, significantly reducing boundary errors. It excels in clearly distinguishing between cells clustered together, presenting a substantial improvement over Model A. These findings underscore the superior performance of our proposed model, emphasizing its potential applicability in diverse image segmentation tasks.



(a) (b) (c)

Fig. 5. **Sample Outputs**: (a) Input image (b) Predicted Segmentation masks using Vanilla Mask RCNN. (c) Predicted Segmentation masks using Mask RCNN that also incorporates our proposed MSFL and CSAM. We observe that the quality of segmentation mask improves and the model is able to better distinguish between overlapping cells and cell borders.

C. Comparison on Model Size

We note that transitioning from ResNet50 (A) to ResNeXt101 (B) as the backbone of the baseline Mask RCNN results in an augmentation of approximately 63 million trainable parameters, with only a marginal increase of 0.84% in mIoU. Conversely, the incorporation of the MSFL technique leads to a minimal increment of 0.5 million trainable parameters in Mask RCNN, accompanied by a notable improvement of 1.86% in mIoU. Subsequently, the addition of CSAM results in a mere increase of 0.1 million parameters over Mask RCNN with MSFL, yet yields an improvement of 0.80% in mIoU as compared to it, underscoring the significance of the CSAM module. This highlights the lightweight nature of the CSAM module alongside its substantial impact on performance.

D. Assessing Model Generalizability with Another Dataset

We train our refined model using a similar approach on another dataset of yeast cell segmentation in microstructures and evaluate its performance using the mean Average Precision@[0.5:0.95], a widely used metric for instance segmentation. Results obtained on this dataset are presented in Table 2. Compared to the baseline vanilla Mask R-CNN, our model demonstrates improved performance on the secondary dataset, achieving a 1.86% increase in mAP@[0.5:0.95]. This validation underscores the robustness and adaptability of our proposed segmentation approach across diverse biological contexts.

Table 2. Experimental Results on Yeast Cell dataset: Comparison of baseline Mask R-CNN with the outcomes achieved with our proposed multiscale feature learning and cross-attention module

Model	Backbone	mAP@[0.5:0.95]
Mask RCNN	ResNeXt101	0.7504
Mask RCNN+MSFL	ResNeXt101	0.7608
Mask RCNN+MSFL+CSAM	ResNeXt101	**0.7690**

V. Conclusion

In our ablation study on the SegPC dataset, we examined the impact of crucial factors, such as model backbone, training augmentations, Multiscale Feature Learning, and the

Cross Spatial Attention Module, on instance segmentation performance. The comprehensive experiments and summarized results underscore the effectiveness of our proposed modifications. Notably, we achieved a substantial improvement in mIOU by incorporating augmentations, utilizing advanced backbones, applying MSFL, and utilizing the lightweight CSAM. Our model demonstrated superior performance compared to the baseline Mask RCNN, exhibiting enhanced segmentation accuracy, smoother boundaries, and an improved ability to distinguish between clustered cells.

In addition to the compelling results on the SegPC dataset, we extended our analysis to a yeast cell segmentation dataset, further validating the versatility of our proposed model. This cross-dataset validation consistently demonstrated improved performance, affirming the adaptability and robustness of our approach across diverse instance segmentation scenarios.

Importantly, the CSAM, as a lightweight module, not only contributed to the observed positive outcomes but also showcased its potential for easy integration into other domain-specific segmentation tasks. These findings emphasize the broader applicability and efficacy of our proposed model, establishing it as a valuable tool for various image segmentation challenges.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Dipankar, Rachita, Isha, and Anubha conceived the presented methodology and planned the experiments. Dipankar, Rachita, Isha, and Anish performed the computations. All authors discussed and interpreted the results and contributed to writing the draft. All authors approved the final version.

Funding

This project was funded by SERB, Department of Science and Technology, Govt. of India through Grant No. SPF/2021/000209.

Acknowledgements

AG would like to thank SERB, Department of Science and Technology, Govt. of India for funding the project through Grant No. SPF/2021/000209. In addition, authors would also like to thank the Centre for Excellence in Healthcare, IIIT-Delhi and the Infosys Centre for AI for providing the compute facility at different stages to complete this work.

References

[1] S. Gehlot, A. Gupta, and R. Gupta, "A CNN-based unified framework utilizing projection loss in unison with label noise handling for multiple myeloma cancer diagnosis," *Medical Image Analysis*, vol. 72, p. 102099, 2021.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, October 5-9, 2015, pp. 234–241.

[3] S. Gehlot, A. Gupta, and R. Gupta, "EdNFC-Net: Convolutional neural network with nested feature concatenation for nuclei-instance segmentation," in *Proc. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1389–1393.

[4] P. Kumar, P. Nagar, C. Arora, and A. Gupta, "U-SegNet: fully convolutional neural network based automated brain tissue segmentation tool," in *Proc. 2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3503–3507.

[5] P. O. Pinheiro, R. Collobert, and P. Dollar, *Learning to Segment Object Candidates*, 2015.

[6] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollar, *Learning to Refine Object Segments*, 2016.

[7] A. Kirillov, E. Mintun, N. Ravi *et al*., *Segment Anything*, 2023.

[8] K. M. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[9] Z. W. Cai and N.Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," arXiv preprint arXiv:1906.09756, 2019.

[10] K. Chen, J. M. Pang *et al*., "Hybrid Task Cascade for Instance Segmentation," arXiv preprint arXiv:1901.07518, 2019.

[11] K. Xu, J. Ba *et al*., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. International Conference on Machine Learning (ICML)*, 2015, pp. 2048–2057.

[12] F. Wang, M. Q. Jiang *et al*., "Residual attention network for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3156–3164.

[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[14] B. Z. Chen, Y. S. Liu, Z. Zhang, G. M. Lu, and A. W. K. Kong, "TransAttUnet: Multi-level attention-guided U-net with transformer for medical image segmentation," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

[15] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," arXiv preprint arXiv:1807.06514, 2018.

[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[17] J. Fu, J. Liu, H. J. Tian, Y. Li, Y. J. Bao, Z. W. Fang, and H. Q. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3146–3154.

[18] X. L. Wang, R. Girshick, A. Gupta, and K. M. He, "Efficient channel attention," in *Proc. European Conference on Computer Vision (ECCV)*, 2020, pp. 679–696.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234–241, 2015.

[20] A. Gupta, S. Gehlot *et al*., "SegPC-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images," *Medical Image Analysis*, vol. 83, p. 102677, 2023.

[21] A. Gupta, R. Gupta, S. Gehlot, and S. Goswami, "Segpc-2021: Segmentation of multiple myeloma plasma cells in microscopic images," *IEEE Dataport*, vol. 1, no. 1, p. 1, 2021.

[22] A. Gupta, P. Mallick, O. Sharma, R. Gupta, and R. Duggal, "PCSeg: Color model driven probabilistic multiphase level set based tool for plasma cell segmentation in multiple myeloma," *PloS one*, vol. 13, no. 12, pp. e0207908, 2018.

[23] P. Rai, S. Gehlot, R. Gupta, and A. Gupta, "LIDACS: A lightweight domain adaptive cell segmentation framework," in *Proc. the 2023 Asia Conference on Artificial Intelligence, Machine Learning and Robotics*, 2023, pp. 1–6.

[24] A. Gupta, R. Duggal, S. Gehlot, R. Gupta, A. Mangal, L. Kumar, Ni. Thakkar, and D. Satpathy, "GCTI-SN: Geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images," *Medical Image Analysis*, vol. 65, pp. 101788, 2020.

[25] C. Reich, T. Prangemeier, A. O. Francani, and H. Koeppl, "An instance segmentation dataset of yeast cells in microstructures," in *Proc. 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, July 2023.