

Article

MC-SlotNet: Multiplicity-Consistent Slot-Based Full-Cell Instance Segmentation for Overlapping Plant Suspension-Culture Microscopy

Touseef Ur Rehman ¹, Saba Latif ², Muhammad Talha Shabbir ³, Meijin Guo ^{1,*}
and Muhammad Rameez Ur Rahman ^{4,*}

¹ State Key Laboratory of Bioreactor Engineering, East China University of Science and Technology, Shanghai 200237, China; touseef@mail.ecust.edu.cn

² Department of Computer Science, Sapienza University of Rome, 00185 Rome, Italy; saba.latif@uniroma1.it

³ Pak-Austria Fachhochschule Institute of Applied Sciences and Technology, Haripur 22620, Pakistan; mt3125682@gmail.com

⁴ Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, 30171 Venezia, Italy

* Correspondence: guo_mj@ecust.edu.cn (M.G.); muhammad.rahman@unive.it (M.R.U.R.)

Abstract

Overlapping cells in plant suspension-culture microscopy pose a particular challenge, for instance, segmentation because a single pixel may belong to more than one cell. Most standard instance-segmentation methods are not designed for this setting and tend to treat overlapping objects as mutually exclusive regions. We instead represent each cell as an independent full-cell instance and introduce MC-SlotNet, an architecture that separates competitive object-slot feature assignment from mask decoding. This allows multiple predicted masks to occupy the same image region. We further introduce a mask-level multiplicity-consistency loss that encourages the predicted number of masks covering a pixel to agree with the underlying cell occupancy. We evaluate MC-SlotNet on a newly annotated dataset of 53 *Siraitia grosvenorii* suspension-culture micrographs containing 4131 full-cell instances acquired at 4×–40× magnification. Using grouped five-fold cross-validation and an overlap-preserving evaluation protocol, we compare the method with Mask R-CNN, SOLOv2, and Mask2Former. MC-SlotNet achieves the best performance on AP50 (0.800), mAP50 : 95 (0.565), F150 (0.842), all-ground-truth Dice (0.761), AJI+ (0.754), overlap-region Dice (0.705), and overlap-instance recall (0.828). Its AP75 (0.649) is comparable to Mask2Former's (0.651). MC-SlotNet also has the lowest inference time among the evaluated methods, at 1.113 s/image. These results indicate that decoding full-cell masks independently, rather than enforcing an exclusive partition of image pixels, is well-suited to instance segmentation in plant suspension-culture microscopy images with substantial cell overlap.



Academic Editor: Hongying Meng

Received: 13 August 2026

Revised: 28 August 2026

Accepted: 1 September 2026

Published: 4 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: instance segmentation; overlapping cell segmentation; plant cell suspension culture; slot attention; microscopy image analysis; multi-instance mask prediction

1. Introduction

Plant biotechnology increasingly relies on controlled culture systems for the production and study of plant-derived compounds. Among these systems, plant cell and suspension cultures provide a controllable platform for producing high-value metabolites, natural products, flavors, pigments, and other bioactive compounds [1,2]. Their growth and

productivity are influenced by several culture characteristics, including cell concentration, aggregation, morphology, viability, and changes in color. Microscopic imaging provides practical means of examining these characteristics and can complement conventional biochemical assays by enabling direct assessment of cell number, size, shape, aggregation, and temporal morphological changes. Accordingly, computer-assisted image analysis has been used to quantify growth and morphology in plant cell suspension cultures [3,4].

Recent advances in deep learning have further expanded the use of automated microscopy analysis for tasks such as object detection and segmentation [5]. Instance segmentation is particularly relevant because it identifies individual objects and assigns a separate pixel-level mask to each detected instance [6]. In biomedical imaging, widely used approaches include U-Net [7], StarDist [8], Cellpose [9], and PlantSeg [10]. These methods have been highly effective for semantic segmentation, shape-constrained detection, or separation of touching objects. However, many commonly used representations ultimately assign each image pixel to a single object or region. This assumption becomes problematic when cells overlap in a two-dimensional microscopy projection, because a shared pixel may legitimately belong to more than one biological instance.

General-purpose instance-segmentation methods provide greater flexibility because they predict separate masks for individual objects. Mask R-CNN [11], for example, predicts an instance mask for each detected region, whereas SOLOv2 [12] generates location-conditioned dynamic kernels over shared mask features. Mask2Former uses learned object queries and masked attention to produce instance-level predictions [13]. Because their predicted masks can be retained independently, these architectures can represent spatial overlap in principle. Nevertheless, reliable reconstruction of heavily overlapping full-cell shapes remains challenging, particularly when multiple cells share substantial portions of their projected area.

Several recent studies have addressed overlapping biological instances more directly. VB-SOLO adapted the SOLOv2 framework for segmentation of overlapping epithelial cells [14]. IAUNet combined U-Net-derived pixel features with instance-aware queries for overlapping microscopic objects [15]. DiffusionSplit approached the problem through stochastic symmetry breaking, allowing multiple pixel-level instances to be recovered without forcing the output into an exclusive label map, although this requires iterative diffusion-based inference [16]. Together, these studies demonstrate the importance of representations that preserve multiple instance memberships in regions of overlap. However, to our knowledge, comparatively little attention has been given to this problem in plant suspension-culture microscopy, where dense aggregation and translucent cell boundaries can produce extensive projected overlap.

In this study, we therefore represent the segmentation target as a set of independent full-cell masks rather than as a mutually exclusive partition of the image. Under this formulation, multiple annotated cell instances may occupy the same pixel. The task is related to amodal segmentation, which seeks to recover the complete extent of partially occluded objects [17]. We use the more conservative term full-cell segmentation, because suspension-culture microscopy does not necessarily provide a well-defined front-to-back depth ordering. Instead, the objective is to reconstruct the annotated apparent two-dimensional outline of each cell, including portions that overlap with neighboring cells, without attempting to infer an unobserved three-dimensional structure.

To address this problem, we propose MC-SlotNet, an overlap-aware full-cell instance-segmentation architecture that separates object-feature assignment from final spatial occupancy. Center-guided object queries are refined through competitive slot attention, encouraging the model to develop distinct object representations. Each refined slot then independently generates a dynamic full-cell mask from shared high-resolution features.

Because the final masks are decoded independently rather than normalized across instances, multiple predicted cells can occupy the same spatial location. In addition, a mask-level multiplicity-consistency objective supervises the aggregate predicted occupancy in regions where several annotated cells overlap.

The main contributions of this work are as follows:

1. We formulate overlapping plant suspension-culture microscopy as a full-cell instance-segmentation problem in which multiple cell instances may legitimately occupy the same projected pixels.
2. We introduce MC-SlotNet, an overlap-aware architecture that combines center-guided object-slot initialization and competitive feature assignment with independently decoded full-cell masks, thereby separating feature assignment from final spatial occupancy.
3. We introduce a mask-level multiplicity-consistency objective that supervises the aggregate occupancy of independently predicted masks in overlapping regions, together with auxiliary center, multiplicity, and boundary supervision.
4. We evaluate the proposed approach on a newly annotated suspension-culture microscopy dataset containing 4131 full-cell instances across four objective magnifications, using grouped five-fold cross-validation and an overlap-preserving evaluation protocol against Mask R-CNN, SOLOv2, and Mask2Former.

2. Materials and Methods

2.1. Plant Cell Culture

In this study, cell suspension cultures of *Siraitia grosvenorii* (*S. grosvenorii*) were established from cotyledons [18]. Approximately 50 g of callus cells were inoculated into 500 mL culture flasks containing 200 mL of Gamborg's B5 (B5) medium. This medium was supplemented with 4 mg/L of 1-naphthylacetic acid (NAA), 0.2 mg/L of 6-benzylaminopurine (6-BA), and 30 g/L of sucrose, with the pH maintained at 5.8. The media were sterilized at 121 °C for 25 min using an autoclave. Cultures were established under sterile conditions and incubated on a rotary shaker at 115 rpm and 25 °C in darkness. For microscopic imaging, the cells were obtained from the cell suspension culture on the seventh day after inoculation.

2.2. Microscopic Imaging and Physical Calibration

Samples were taken from the suspension culture, placed on glass slides, and examined under a microscope and imaged with a calibrated MShot image-110 analysis system (Micro-shot Technology Co., Ltd., Guangzhou, China). The image set contained RGB TIFF files with native dimensions of 1536×2040 or 3072×4088 pixels. Images were acquired at $4\times$, $10\times$, $20\times$, and $40\times$ objective magnification.

Spatial calibration was obtained from calibration images and expressed in pixels per micrometer (Table 1). The $10\times$ calibration, 2.437887 pixels/ μm , was selected as the common processing scale. An image acquired at scale s_i was resampled by $2.437887/s_i$ before crop extraction and model inference. This normalization reduces the apparent cell-size shift among magnifications while preserving the original held-out image as the independent evaluation unit. Figure 1 summarizes the workflow used to obtain and prepare the microscopy dataset, from suspension-culture sampling and slide preparation through image acquisition, manual annotation, and COCO-format conversion.

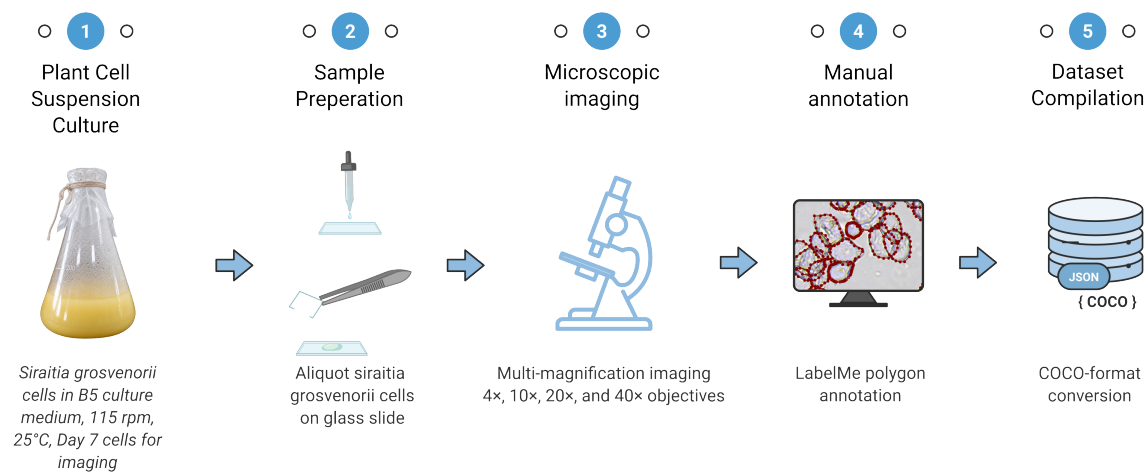


Figure 1. Workflow for microscopy image acquisition, annotation, and dataset construction. *Siraitia grosvenorii* suspension-culture cells were maintained in B5 medium and sampled on day 7. An aliquot was mounted on a glass slide and examined using the MShot digital microscopy system at 4×, 10×, 20×, and 40× objective magnifications. Individual cells were manually outlined in LabelMe, with overlapping full-cell instances retained as independent polygons, before conversion to the COCO annotation format used for model development.

Table 1. Physical calibration used for scale normalization.

Objective	Pixels/μm	Linear Scale to 10×
4×	1.083715	2.2496
10×	2.437887	1.0000
20×	4.854299	0.5022
40×	9.842450	0.2477

2.3. Annotation and Full-Cell Target Representation

We manually annotated all microscopic images in LabelMe [19]. Each cell was traced independently according to its apparent full two-dimensional outline, including regions shared with neighbouring cells. Overlapping cells were therefore not merged or assigned mutually exclusive pixels. Annotation was performed by a multidisciplinary team. A biotechnology/bioengineering domain expert first traced each cell’s apparent full two-dimensional outline, drawing on knowledge of *Siraitia grosvenorii* cell morphology to resolve ambiguous cases. Each image was then independently cross-checked by a second team member with machine learning and data-mining expertise, who reviewed polygon consistency, class labels, and suitability for the COCO-style instance representation used for training. Discrepancies identified during cross-checking were resolved by joint discussion between the domain expert and the reviewing annotator before an image was finalized. To avoid speculative boundaries, cells with unclear extents—such as those nearly fully occluded by neighboring cells or debris—were excluded from the annotated dataset. However, overlapping cells were retained as independent full-cell instances as long as their shared boundaries remained visually resolvable. LabelMe annotations were converted to a COCO-compatible instance representation [20], with each polygon retained as an independent cell instance. Image dimensions, polygon validity, class labels, and associated physical metadata were checked during conversion.

For an image domain Ω of height H and width W , the annotation was represented as

$$\mathcal{G} = \left\{ G_i \in \{0, 1\}^{H \times W} \right\}_{i=1}^N, \quad (1)$$

where $G_i(p) = 1$ indicates that pixel p belongs to cell i . Because the instance masks are independent, $G_i(p) = G_j(p) = 1$ is permitted for $i \neq j$. The exact pixel-wise multiplicity was

$$K_{\text{exact}}(p) = \sum_{i=1}^N G_i(p), \quad p \in \Omega. \quad (2)$$

For auxiliary multiplicity classification, this quantity was clipped as

$$K_{\text{aux}}(p) = \min(3, K_{\text{exact}}(p)), \quad (3)$$

yielding classes background, one cell, two cells, and three-or-more cells. This grouping reduces sparsity among high-order overlap classes while the independent instance masks and mask-level multiplicity-consistency target retain the observed higher-order occupancy, including multiplicities above three. The original independent masks and exact multiplicity were retained unchanged.

2.4. Dataset Composition

The full dataset contained 53 microscopy images and 4131 annotated full-cell instances. Detailed dataset characteristics are shown in Table 2. All images contained overlapping foreground with a mean image-level overlap fraction of 16.02% and the foreground-area-weighted overlap fraction of 16.9%. The maximum observed multiplicity was five cells at a single pixel, confirming that overlap was a frequent rather than exceptional property of the dataset.

To illustrate annotation quality, annotation examples spanning all four magnifications are shown in Figure 2. The examples include isolated cells, dense aggregates, severe overlap, and cells intersecting the image boundary. Each apparent full-cell outline was retained as an independent instance even when portions of two or more polygons occupied the same projected pixels.

Table 2. Composition of the annotated microscopy dataset by magnification.

Magnification	Images	Cell Instances	Mean Instances/Image	Instance Range	Pixels/ μm
4×	6	1688	281.3	127–833	1.084
10×	16	1279	79.9	38–238	2.438
20×	13	767	59.0	19–133	4.854
40×	18	397	22.1	5–38	9.842
Total	53	4131	77.8	5–833	–

2.5. Cross-Validation and Experimental Governance

We use a five-fold grouped cross-validation scheme across all 53 images. The detailed composition is shown in Table 3. Folds were constructed to balance magnification as far as possible while keeping all instances and crops from the same image within a single fold. Due to practical constraints in maintaining image-level grouping, fold sizes varied from 9 to 12 images. The same fold assignment was used for every method.

For each fold, the held-out images were excluded from optimization and loss computation, and all models were trained for a fixed 80 epochs rather than selecting checkpoints from held-out performance.

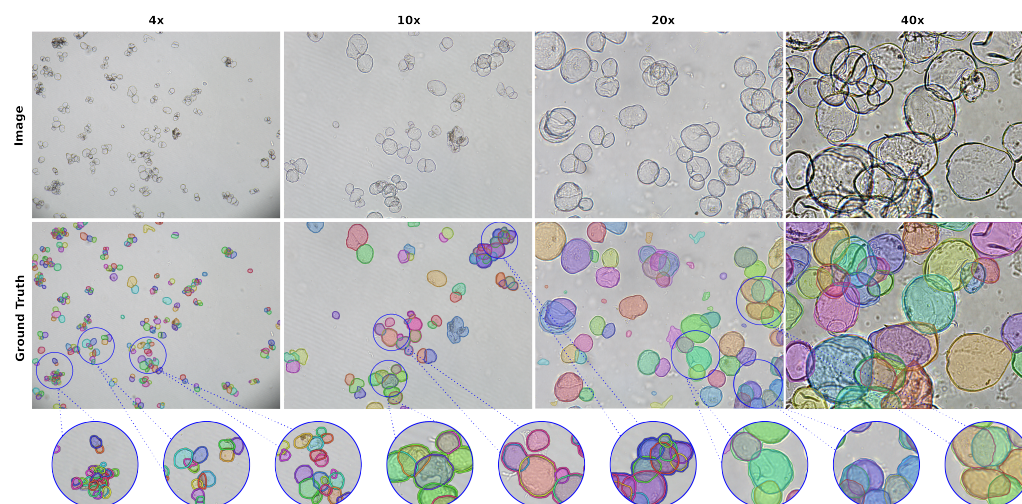


Figure 2. Example microscopy images and independent full-cell annotations across 40 $\times$, 20 $\times$, 10 $\times$, and 4 $\times$ magnification. Examples include isolated cells, crowded aggregates, substantial overlap, and border-intersecting instances. Each color denotes an independently annotated full-cell mask; shared image pixels may therefore belong to more than one cell.

Table 3. Composition of the five grouped cross-validation folds, including per-fold overlap statistics.

Fold	Images	Instances	4 $\times$	10 $\times$	20 $\times$	40 $\times$	Mean Overlap %	Median Overlap %	Max Overlap %
0	12	1043	2	3	3	4	15.8	14.2	41.3
1	9	451	1	3	2	3	16.5	15.0	39.8
2	12	839	1	4	3	4	15.1	13.6	42.7
3	10	616	1	3	3	3	17.2	15.8	44.1
4	10	1182	1	3	2	4	16.4	14.9	40.5
Total	53	4131	6	16	13	18	16.2	14.7	44.1

2.6. Image Preprocessing and Training-Crop Sampling

We resample images to the common physical scale before training. Each channel was normalized using its 1st and 99th intensity percentiles and clipped to [0, 1]; inputs to ImageNet-pretrained encoders were subsequently standardized using the ImageNet channel statistics.

Training used 512 $\times$ 512-pixel crops. Ninety percent of crops were centered on annotated cells, with the remainder sampled more generally to retain background and density variation. Instances whose complete bounding boxes crossed a crop boundary were excluded as positive mask targets, and truncated regions together with a 64-pixel crop-border margin were excluded from dense losses.

Geometric augmentation comprised horizontal and vertical flips and 90° rotations. Photometric augmentation included mild brightness, contrast, gamma, Gaussian-noise, and Gaussian-blur perturbations.

2.7. Synthetic Overlap Augmentation

Instance-level copy–paste augmentation was used [21] to increase the diversity of overlap configurations observed during training. Low-overlap source cells were drawn exclusively from the current training folds, transformed by random rotation and 0.85–1.15 scaling, and inserted into another training crop with a target overlap fraction of approximately 0.05–0.55. Soft blending and alpha values of 0.70–0.95 reduced visible paste boundaries.

Synthetic overlap was applied with a probability of 0.20, with at most one pasted cell per crop. Original and pasted masks remained separate, preserving independent full-cell supervision and exact overlap targets. Cells from the held-out fold were never used as copy-paste sources.

2.8. Multiplicity-Consistent Slot Network Architecture

We designed MC-SlotNet for full-cell instance segmentation in microscopy, where independently annotated cells may legitimately occupy the same image pixels. The architecture of MC-SlotNet is shown in Figure 3. The model combines an ImageNet-pretrained Swin-Tiny encoder [22], a high-resolution U-Net-style decoder [7], center-guided object-slot initialization, competitive slot refinement, and instance-conditioned dynamic mask decoding. The slot-refinement mechanism is related to object-centric slot attention [23], while the dynamic masks follow the general instance-conditioned prediction principle used in CondInst and SOLOv2 [12,24].

The central design principle is to separate *object-feature assignment* from *spatial occupancy*. During slot refinement, decoder features are assigned competitively across object slots. Final instance masks, however, are decoded independently and are therefore not constrained to form a mutually exclusive partition. Multiple predicted full-cell masks may consequently assign positive probability to the same spatial location. A mask-multiplicity consistency objective constrains the aggregate occupancy of these independent masks using the annotated overlap structure.

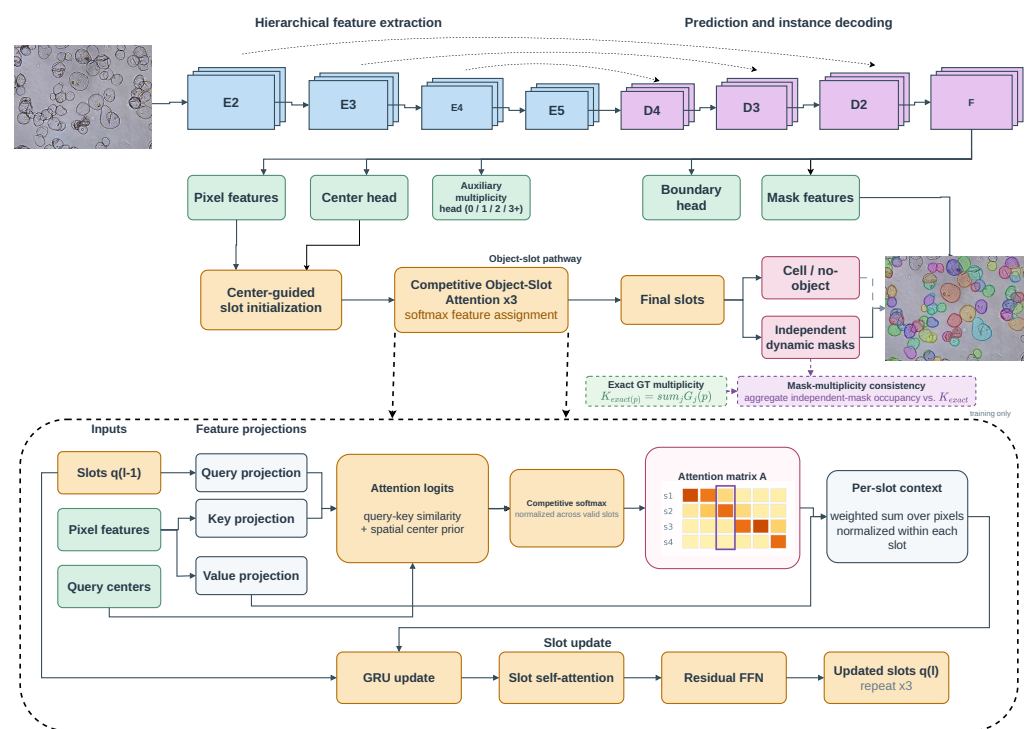


Figure 3. Architecture of MC-SlotNet. An ImageNet-pretrained Swin-Tiny encoder provides four hierarchical feature maps to a stride-4 decoder. Center candidates initialize object slots from local decoder features, normalized position, center confidence, and learned query embeddings. Three competitive slot-refinement layers construct object-specific representations. Final slots produce cell/no-object scores and dynamic parameters that operate on a shared mask-feature map to generate independently decoded full-cell masks. These masks are not mutually exclusive and may overlap spatially. Dense center, multiplicity, and boundary heads provide auxiliary supervision, while mask-level multiplicity consistency constrains the aggregate occupancy of the final instance set.

2.8.1. Hierarchical Encoder and Pixel Decoder

The MC-SlotNet used an ImageNet-pretrained Swin-Tiny encoder [22]. Encoder features E_2 , E_3 , E_4 , and E_5 were extracted at spatial strides 4, 8, 16, and 32, with 96, 192, 384, and 768 channels, respectively. Before encoding, percentile-normalized RGB inputs were standardized using the ImageNet channel means (0.485, 0.456, 0.406) and standard deviations (0.229, 0.224, 0.225).

A top-down skip-connected decoder bilinearly upsampled the deepest features, concatenated the corresponding encoder skip features, and applied convolutional refinement. The three decoder stages produced 256, 192, and 128 channels at strides 16, 8, and 4, respectively. The decoder-stage subscripts D_4 , D_3 , and D_2 denote stages rather than spatial strides. A final refinement block preserved the stride-4 resolution and produced decoder feature tensor F .

Two projections of F were used by the object-slot pathway: a 128-channel pixel-feature map F^{pix} and a 64-channel shared mask-feature map F^{mask} . Three additional stride-4 heads predicted the center heatmap H^{ctr} , auxiliary multiplicity logits H^{mult} , and foreground-union boundary logits H^{bnd} . The multiplicity head provides auxiliary overlap-aware supervision only; it does not condition the competitive attention or the dynamic mask decoder.

2.8.2. Center-Guided Slot Initialization

We obtain the potential object centers by applying a 3×3 local-maximum operator to $\sigma(H^{\text{ctr}})$ and ranking the resulting peaks by center confidence, following the center-point representation of CenterNet [25]. At most $Q = 96$ object queries were used for a 512×512 normalized crop. In the general case, if the number of complete target instances were to exceed Q , the model would retain at most 96 query assignments for instance-level matching, while all annotated cells would continue to contribute to the dense center, multiplicity, and boundary supervision. This fallback was not invoked in the present experiments. For candidate center $c_i = (x_i, y_i)$ with center confidence s_i , the initial slot was

$$q_i^{(0)} = W_{\text{pix}} F^{\text{pix}}(c_i) + \text{MLP}_{\text{pos}} \left[\frac{x_i}{W_4 - 1}, \frac{y_i}{H_4 - 1}, s_i \right] + e_i, \quad (4)$$

where H_4 and W_4 denote the stride-4 feature dimensions and $e_i \in \mathbb{R}^{128}$ is a learned embedding associated with the i th position in the confidence-ranked candidate set. Thus, the initial representation combines local appearance, spatial location, center confidence, and a learned query embedding.

During training, ground-truth interior centers were injected into the query set with a probability of 0.75. Injected centers were perturbed by Gaussian noise with a standard deviation of one stride-4 feature pixel, and predicted center candidates within 1.5 feature pixels of an injected center were excluded. Remaining query positions were filled by predicted center peaks. This partial teacher forcing stabilized early optimization while retaining exposure to predicted-center errors. All constructed training queries participated in matching; unmatched queries were supervised as *no object*.

Center extraction and confidence ranking are discrete operations. Gradients were therefore not propagated through the selected peak indices or their ranking. The selected center confidence remained an input to Equation (4) and could receive gradients through the downstream slot pathway. At inference, query validity was determined from the predicted center candidates; the corresponding confidence threshold and minimum-query rule are reported in Section 2.10.

2.8.3. Competitive Object-Slot Attention

Let Ω_4 denote the stride-4 decoder grid. For slot i and location $r \in \Omega_4$, the attention logit was

$$\ell_{ir} = \frac{(W_q q_i)^T (W_k f_r)}{\sqrt{d}} - \frac{\|u_r - \bar{c}_i\|_2^2}{2\sigma_{\text{att}}^2}, \quad (5)$$

where f_r is the pixel feature at location r , $u_r \in [0, 1]^2$ is its normalized coordinate, $\bar{c}_i \in [0, 1]^2$ is the normalized query-center coordinate, and $\sigma_{\text{att}} = 0.30$ is the spatial attention radius. No multiplicity-dependent term is included in the attention logits.

Let $\mathcal{V} = \{i : v_i = 1\}$ denote the set of valid queries. Competitive attention is defined by a softmax over this set:

$$A_{ir} = \begin{cases} \frac{\exp(\ell_{ir})}{\sum_{j \in \mathcal{V}} \exp(\ell_{jr})}, & i \in \mathcal{V}, \\ 0, & i \notin \mathcal{V}. \end{cases} \quad (6)$$

Hence, $\sum_{i=1}^Q A_{ir} = 1$ whenever at least one valid query is present. This normalization creates competition among object slots for decoder features, but A_{ir} is used only for feature aggregation and is not interpreted as an instance-segmentation probability (Figure 4). At inference, at least the eight highest-scoring center candidates were retained regardless of the nominal center threshold. Consequently, the valid query set $\mathcal{V}$ was non-empty during normal inference.

The context vector for slot i was

$$z_i = \frac{\sum_{r \in \Omega_4} A_{ir} W_v f_r}{\sum_{r \in \Omega_4} A_{ir} + \varepsilon}. \quad (7)$$

Each refinement layer incorporated this context through a gated recurrent update, followed by four-head slot self-attention and a residual feed-forward block [26]. Three refinement layers were stacked, with slot dimension 128. The query-center coordinates remained fixed throughout slot refinement; refinement updated the slot representation but did not regress the associated spatial center.

MC-SlotNet decouples competitive object assignment from non-exclusive mask occupancy

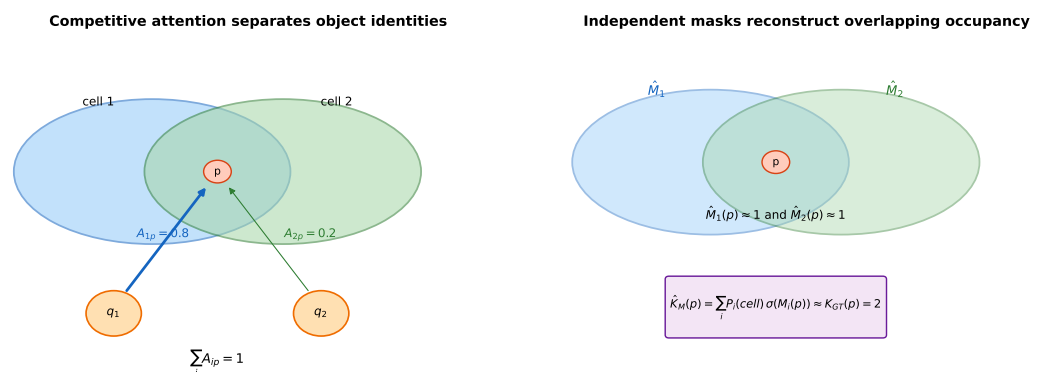


Figure 4. Decoupling competitive feature assignment from overlapping mask reconstruction. Slot attention is normalized across valid queries at each decoder location and is used to construct distinct object representations. Final masks are decoded independently and are not mutually exclusive. Their aggregate occupancy is constrained by the corresponding ground-truth mask multiplicity.

For variant A4 only, the predicted local cell multiplicity was also used to modulate slot attention. The auxiliary multiplicity probabilities were converted to an expected count,

$$\hat{K}_{\text{dense}}^{(4)}(r) = \sum_{k=0}^{K_{\max}} k P(K_{\text{aux}}^{(4)}(r) = k \mid I),$$

and the attention logit became

$$\ell_{ir}^{\text{A4}} = \ell_{ir} + 0.75 \frac{\hat{K}_{\text{dense}}^{(4)}(r)}{K_{\max}}. \quad (8)$$

A4 used independent sigmoid attention, $A_{ir}^{\text{A4}} = v_i \sigma(\ell_{ir}^{\text{A4}})$. Thus, locations predicted to contain several overlapping cells were allowed to receive stronger responses from multiple slots. We use this multiplicity modulation ablation A4.

2.8.4. Independent Dynamic Full-Cell Masks

Each refined slot produced two classification logits corresponding to *cell* and *no object*, together with dynamic mask parameters $\theta_i = [w_i^{\text{feat}}, a_i^x, a_i^y, a_i^r, b_i]$, where $w_i^{\text{feat}} \in \mathbb{R}^{64}$ acts on the shared mask-feature map. For stride-4 location r , let $\Delta_i^x(r) = u_r^x - \tilde{c}_i^x$ and $\Delta_i^y(r) = u_r^y - \tilde{c}_i^y$. The dynamic mask logit was

$$M_i^{(4)}(r) = \frac{(w_i^{\text{feat}})^T F^{\text{mask}}(r)}{\sqrt{64}} + a_i^x \Delta_i^x(r) + a_i^y \Delta_i^y(r) + a_i^r \left[(\Delta_i^x(r))^2 + (\Delta_i^y(r))^2 \right] + b_i. \quad (9)$$

The corresponding soft mask is $\hat{G}_i^{(4)}(r) = \sigma(M_i^{(4)}(r))$. This lightweight parameterization uses the shared mask-feature tensor to represent non-linear cell shape, while the first-order and radial second-order coordinate terms provide an explicit query-relative spatial prior.

Mask matching and mask-related training losses were evaluated on the stride-4 domain. Ground-truth instance masks were resized to the same resolution using nearest-neighbor interpolation. During inference, predicted mask logits were bilinearly upsampled to the input-image resolution before sigmoid conversion, thresholding, area filtering, and full-image merging. Because each query is decoded independently, the resulting full-cell masks are not constrained to be mutually exclusive and may overlap spatially.

2.8.5. Target Construction and Hungarian Assignment

The center target contained Gaussian peaks at the annotated interior centers, with radii derived from the stride-4 instance bounding-box dimensions as in CenterNet [25].

For the auxiliary multiplicity classifier, the full-resolution multiplicity map was converted to a stride-4 four-class target using maximum aggregation:

$$K_{\text{aux}}^{(4)}(r) = \min \left[K_{\max}, \max_{p \in \mathcal{B}(r)} K_{\text{exact}}(p) \right], \quad K_{\max} = 3, \quad (10)$$

where $\mathcal{B}(r) \subset \Omega$ denotes the full-resolution region associated with decoder location r . Maximum aggregation was used because narrow overlap interfaces may occupy only a small fraction of a stride-4 bin and could otherwise disappear during downsampling. Accordingly, $K_{\text{aux}}^{(4)}$ represents the maximum local occupancy within each decoder bin rather than an area-averaged multiplicity. Values above three are grouped into the 3+ auxiliary class.

The boundary target was constructed from the foreground union $U(p) = \mathbb{I}[K_{\text{exact}}(p) > 0]$. A two-pixel morphological boundary of this union was computed at the input resolution and projected to the stride-4 dense-head resolution using maximum aggregation.

A separate, uncapped target was used for mask-multiplicity consistency. Each ground-truth instance mask was resized independently to the stride-4 grid using nearest-neighbor interpolation, $G_j^{(4)} = \mathcal{N}_{1 \rightarrow 4}(G_j)$, and the target multiplicity was

$$K_{\text{mask}}^{(4)}(r) = \sum_{j=1}^N G_j^{(4)}(r). \quad (11)$$

Unlike the auxiliary target, $K_{\text{mask}}^{(4)}$ is not clipped and therefore retains higher-order occupancy. Because the instance masks are resampled independently before summation, $K_{\text{mask}}^{(4)}$ represents the multiplicity of the stride-4 resampled instance masks and is not, in general, identical to a direct downsampling of the full-resolution multiplicity map K_{exact} .

Predicted queries and ground-truth instances were matched one-to-one with the Hungarian algorithm. The optimal injective assignment was

$$\mu^* = \arg \min_{\mu} \sum_{j=1}^N \mathcal{C}_{\mu(j),j}, \quad \mathcal{C}_{ij} = 2\mathcal{C}_{ij}^{\text{cls}} + 2\mathcal{C}_{ij}^{\text{BCE}} + 3\mathcal{C}_{ij}^{\text{Dice}}, \quad (12)$$

where $\mathcal{C}_{ij}^{\text{cls}} = -\pi_i^{\text{cell}}$ and π_i^{cell} is the softmax probability assigned to the cell class. The BCE and Dice components were evaluated between $\hat{G}_i^{(4)}$ and $G_j^{(4)}$ over crop-valid decoder locations. Matched queries were supervised as cells, whereas unmatched training queries were assigned the no-object class with a weight of 0.10. Mask BCE and Dice losses were applied only to matched query–instance pairs.

2.8.6. Training Objective

The training objective combined dense auxiliary supervision, matched-instance supervision, and mask-level multiplicity consistency:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{mult}} + \mathcal{L}_{\text{bnd}}}_{\mathcal{L}_{\text{dense}}} + \underbrace{\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mask,bce}} + \mathcal{L}_{\text{mask,dice}}}_{\mathcal{L}_{\text{inst}}} + \mathcal{L}_{\text{mask-mult}} \quad (13)$$

The center head used a CenterNet-style focal loss, the auxiliary multiplicity head used class-weighted cross-entropy, and the boundary head used BCE and Dice loss. Matched-instance masks were supervised using BCE and Dice loss. To emphasize crowded regions, mask BCE was weighted threefold at stride-4 locations with $K_{\text{aux}}^{(4)} > 1$.

For the mask-multiplicity term, the aggregate predicted occupancy was

$$\hat{K}_M^{(4)}(r) = \sum_{i=1}^Q v_i \pi_i^{\text{cell}} \hat{G}_i^{(4)}(r), \quad (14)$$

where $v_i \in \{0, 1\}$ denotes query validity, π_i^{cell} is the predicted probability that query i corresponds to a cell, $\hat{G}_i^{(4)}(r)$ is the independently decoded stride-4 mask probability at location r , and $d(r) \in \{0, 1\}$ denotes the crop-validity mask. The independently decoded masks allow $\hat{K}_M^{(4)}(r) > 1$. Consistency with the uncapped target $K_{\text{mask}}^{(4)}$ was enforced using a weighted Smooth- L_1 loss:

$$\mathcal{L}_{\text{mask-mult}} = \frac{\sum_r d(r) w_K(r) \text{SmoothL1}(\hat{K}_M^{(4)}(r), K_{\text{mask}}^{(4)}(r))}{\max(1, \sum_r d(r) w_K(r))}, \quad (15)$$

with

$$w_K(r) = 0.25 + 0.75 \mathbb{I}[K_{\text{mask}}^{(4)}(r) > 0] + \mathbb{I}[K_{\text{mask}}^{(4)}(r) > 1].$$

Here, $K_{\text{mask}}^{(4)}$ is the uncapped stride-4 ground-truth multiplicity obtained by summing the independently resampled instance masks, and $w_K(r)$ assigns additional weight to cellular and overlapping locations. Overlap is represented by the independently decoded masks and regularized through $\mathcal{L}_{\text{mask-mult}}$.

The ablation A4 also included an attention-budget loss. Its purpose was to make the combined attention of predicted cell slots reflect the number of cells expected at each location. The attention-derived occupancy was

$$\hat{K}_A^{(4)}(r) = \sum_{i=1}^Q v_i \pi_i^{\text{cell}} A_{ir}^{\text{A4}}, \quad (16)$$

and was compared with the auxiliary multiplicity target using the same overlap-weighted Smooth- L_1 formulation used for multiplicity consistency:

$$\mathcal{L}_{\text{att-budget}} = \text{SmoothL1}(\hat{K}_A^{(4)}, K_{\text{aux}}^{(4)}). \quad (17)$$

In the proposed A1 model, multiplicity supervision is instead applied to the final independent masks and does not constrain attention.

2.9. MC-SlotNet Implementation Details

The primary configuration used an ImageNet-pretrained Swin-Tiny encoder [22], approximately 34.2 million trainable parameters, a 128-channel stride-4 decoder, 128-dimensional object slots, three competitive slot-refinement layers, four self-attention heads, and a maximum of 96 queries per tile. Competitive softmax attention was used for object-feature assignment during slot refinement, whereas the final full-cell masks were decoded independently so that overlapping spatial occupancy remained possible. Pixel-wise multiplicity was predicted by an auxiliary dense head, while an additional mask-multiplicity-consistency objective constrained the aggregate occupancy of the independently decoded masks using the unclipped ground-truth multiplicity. Table 4 lists the configuration used for the experiments.

Table 4. Primary MC-SlotNet implementation configuration.

Component	Configuration
Encoder	ImageNet-pretrained Swin-Tiny [22]; strides 4, 8, 16, and 32; channels 96, 192, 384, and 768
Decoder	Skip-connected top-down decoder; intermediate widths 256, 192, and 128; output stride 4
Pixel and mask features	128-channel pixel features and 64-channel shared mask features
Object slots	Dimension 128; three competitive refinement layers; four self-attention heads; maximum 96 queries per tile
Slot attention	Competitive softmax attention across valid object slots; spatial radius 0.30; no multiplicity-dependent attention bias
Multiplicity classes	Auxiliary dense prediction of background, one cell, two cells, and three or more cells (0, 1, 2, 3+); auxiliary target clipped at $K_{\text{max}} = 3$
Mask-multiplicity consistency	Aggregate occupancy of the independently decoded masks supervised against the unclipped ground-truth multiplicity
Training query injection	Ground-truth-center injection probability 0.75; Gaussian center jitter $\sigma = 1$ stride-4 pixel
Input crop	512×512 pixels after physical-scale normalization
Training schedule	80 epochs; 256 sampled crops per epoch; batch size 2

Table 4. *Cont.*

Component	Configuration
Optimizer	AdamW; learning rate 1.5×10^{-4} ; weight decay 10^{-4}
Learning-rate schedule	Three-epoch linear warm-up followed by cosine decay
Regularization	Gradient-norm clipping at 5.0; dropout 0
Instance filtering	Score threshold 0.25; mask threshold 0.50; minimum area 30 normalized pixels
Within-tile duplicate rule	Mask IoU at least 0.95 and center distance at most 3 pixels
Full-image inference	512×512 tiles; 128-pixel tile overlap; inference batch size 1

All experiments used the same five grouped cross-validation folds, percentile intensity normalization, physical-scale normalization, center-biased crop sampling, crop-validity masks, and synthetic-overlap augmentation as the comparison methods. Images were normalized to $2.437887 \text{ pixels}/\mu\text{m}$. Synthetic overlap was applied with probability 0.20, with at most one pasted cell per crop. Partially truncated instances were excluded as positive crop targets, and invalid border regions were excluded from dense losses and matching.

Training used AdamW for 80 epochs, with 256 sampled crops per epoch, batch size 2, learning rate 1.5×10^{-4} , weight decay 10^{-4} , a three-epoch linear warm-up followed by cosine decay, and maximum gradient norm 5.0. Mixed precision was disabled. Training and inference were performed on an NVIDIA A40 GPU with 48 GB memory.

2.10. MC-SlotNet Inference and Duplicate Control

At inference, center candidates with a heatmap probability of at least 0.03 were retained, with a minimum of the eight highest-scoring candidates. Preliminary query confidence was

$$s_i^{\text{pre}} = \pi_i^{\text{cell}} (0.5 + 0.5s_i^{\text{ctr}}), \quad (18)$$

and queries with $s_i^{\text{pre}} < 0.25$ were discarded. Mask logits were upsampled to the input resolution and thresholded at 0.50; masks smaller than 30 normalized pixels were removed. Final confidence additionally included the mean mask probability within the retained mask.

Full images were processed using 512×512 tiles with 128-pixel overlap. We did not use conventional overlap-only NMS because high mask overlap may correspond to distinct cells. A lower-scoring prediction was suppressed only when mask IoU was at least 0.95 and center distance was at most three normalized-image pixels. The same overlap-safe principle was used when merging predictions across neighboring tiles.

We fixed all inference thresholds and duplicate-control parameters before the five-fold experiments using a small pilot set of training-only images, then applied them unchanged to all held-out folds. The comparison methods followed the same procedure, with their thresholds and duplicate-suppression settings selected on the same pilot set and kept fixed throughout cross-validation.

2.11. Ablation Variants

We performed ablation experiments to examine the interaction between competitive feature assignment and mask-level multiplicity consistency. To do this, we evaluate five variants. A0 was the competitive baseline, using softmax slot attention without mask-multiplicity consistency. A1 was the proposed formulation, combining competitive softmax attention with mask-multiplicity consistency. A2 replaced competitive softmax with independent sigmoid attention and omitted mask-multiplicity consistency, whereas A3 combined sigmoid attention with the same mask-level consistency objective used by the proposed model. Finally, A4 reproduced the earlier multiplicity-guided non-exclusive

formulation using sigmoid attention, mask-multiplicity consistency, and two additional attention-level mechanisms: multiplicity-guided attention and attention-budget supervision. In A4, the auxiliary multiplicity prediction increased sigmoid-attention responses at locations predicted to contain multiple cells, while the attention-budget loss constrained the aggregate attention occupancy to reflect the local multiplicity target. These two mechanisms were used only in A4 and are defined in Sections 2.8.3 and 2.8.6. All other training and inference settings were kept unchanged across these comparisons.

2.12. Comparison Models

We compared MC-SlotNet with SOLOv2 [12], Mask R-CNN [11], and Mask2Former [13]; Table 5 summarizes how each represents overlapping full-cell instances. All comparison methods were fine-tuned from COCO-pretrained weights using the same five grouped cross-validation folds, 512×512 crops with center-biased sampling, identical augmentation, and the same physical-scale normalization as MC-SlotNet, for 80 epochs on the same NVIDIA A40 GPU. At inference, all methods used identical 512×512 tiling, a minimum instance area of 30 normalized pixels, and a shared duplicate-control rule during tile merging that suppressed a prediction only if confirmed as a near-identical duplicate (mask IoU ≥ 0.94 – 0.95 and center distance within a small fraction of tile size), preserving overlapping but spatially distinct cells.

Table 5. Comparison methods and their representation of overlapping full-cell instances.

Method	Instance Mechanism	Overlap Representation	Role
SOLOv2 [12]	Spatial-grid assignment with dynamically generated mask kernels	Independent instance masks without exclusive pixel assignment	One-stage dynamic-mask baseline
Mask R-CNN [11]	Region proposals with ROI-specific masks	Independent ROI masks	Two-stage detection baseline
Mask2Former [13]	Learned object queries with masked attention	Independent sigmoid query masks retained without panoptic flattening	Transformer set-prediction baseline
MC-SlotNet	Center-guided competitive slots with dynamic full-cell masks	Independent masks with mask-level multiplicity consistency	Proposed overlap-aware model

SOLOv2 used a ResNet-50-FPN backbone (MMDetection 3.3.0), assigning instances to spatial grid locations via dynamic convolution kernels applied to shared mask features; predicted masks were retained independently rather than merged into a mutually exclusive label map. The model was trained using SGD optimizer with momentum 0.9, learning rate 5×10^{-3} and a weight decay of 10^{-4} . Inference used score threshold 0.20 and mask threshold 0.50.

Mask R-CNN used the TorchVision ResNet-50-FPN-v2 implementation, with the top three backbone stages trainable. The model was trained using SGD optimizer with a momentum of 0.9, a learning rate of 5×10^{-3} and a weight decay of 10^{-4} . Inference used a score threshold of 0.20, mask threshold of 0.50, model-level box NMS at IoU 0.95, and a cap of 500 detections per image.

Mask2Former used the COCO-pretrained checkpoint, adapted to a single foreground cell class. The model was trained using AdamW optimizer with the learning rate 5×10^{-5} and a weight decay of 10^{-4} . Inference used a score threshold of 0.20, a mask threshold of

0.50, and a model-level duplicate check ($\text{IoU} \geq 0.94$, center distance ≤ 4 px) prior to the shared tile-merging rule above.

2.13. Evaluation Metrics

All metrics were computed from out-of-fold predictions across the five grouped cross-validation folds. Pooled detection metrics and image-level segmentation metrics were aggregated differently. AP_{50} , AP_{75} , and $\text{mAP}_{50:95}$ were calculated from the pooled set of detections and ground-truth instances across all held-out images. In contrast, F1, all-ground-truth Dice, AJI+, overlap-region Dice, overlap-instance recall, FNR_o , and runtime were calculated per held-out image and then averaged across the out-of-fold image set.

Fixed-threshold instance metrics used one-to-one Hungarian matching based on pairwise mask IoU. Precision, recall, and F1 were reported at IoU 0.50. Average precision was evaluated separately from confidence-ranked predictions at the IoU thresholds 0.50, 0.55, ..., 0.95 using 101-point interpolated precision. AP_{50} , AP_{75} , and $\text{mAP}_{50:95}$ therefore follow the IoU-threshold and interpolation convention of COCO [20].

To account for missed cells, all-ground-truth Dice assigned zero to any ground-truth instance without a match at $\text{IoU} \geq 0.50$:

$$D_{\text{all}} = \frac{1}{N} \sum_{j=1}^N \mathbb{I}[\text{IoU}(P_{\mu(j)}, G_j) \geq 0.50] D(P_{\mu(j)}, G_j). \quad (19)$$

AJI+ additionally aggregates intersections and unions over one-to-one matched instances while penalizing unmatched predictions and ground-truth objects [27,28].

For each ground-truth instance, the overlapping region was

$$O_j = G_j \cap \left(\bigcup_{k \neq j} G_k \right). \quad (20)$$

Overlap-region Dice was evaluated for ground-truth cells with $|O_j| > 0$, assigning zero to missed instances. Overlap-instance recall was the fraction of these overlapping cells matched at $\text{IoU} \geq 0.50$. FNR_o at IoU 0.70 measured the fraction of ground-truth instances not recovered at the stricter IoU threshold.

3. Results

3.1. Primary Five-Fold Setting

Table 6 summarizes detection and full-cell segmentation performance. MC-SlotNet achieved the highest AP_{50} (0.800), $\text{mAP}_{50:95}$ (0.565), F1_{50} (0.842), all-ground-truth Dice (0.761), and AJI+ (0.754). Mask2Former obtained the highest AP_{75} at 0.651, compared with 0.649 for MC-SlotNet. SOLOv2 and Mask R-CNN achieved lower scores than MC-SlotNet and Mask2Former on most metrics. The difference between MC-SlotNet and Mask2Former was small under strict localization: AP_{75} differed by only 0.002. In contrast, MC-SlotNet showed larger advantages in F1 and all-ground-truth Dice, indicating strong recovery of the complete instance set while maintaining accurate full-cell masks. Across folds, MC-SlotNet achieved the highest pooled AP_{50} (0.800), $\text{mAP}_{50:95}$ (0.565), F1_{50} (0.842), all-ground-truth Dice (0.761), and AJI+ (0.754), and these advantages were consistent in direction across all five folds (Table 6). Mask2Former obtained a marginally higher pooled AP_{75} (0.651 versus 0.649). A paired comparison across the five folds gave a mean AP_{75} difference of -0.002 in favor of Mask2Former, with 95% CI $[-0.021, 0.017]$, paired $t(4) = -0.29$, and $p = 0.79$, indicating that this difference is not statistically significant given the available folds. We therefore describe MC-SlotNet and Mask2Former as statistically indistinguishable on AP_{75} ,

while noting that MC-SlotNet’s advantages on the remaining metrics were consistent across all five folds.

Table 6. Per-fold segmentation performance across the five grouped cross-validation folds. “Pooled” is computed from all out-of-fold predictions combined before metric computation; “Mean $\pm$ SD” is the unweighted average $\pm$ standard deviation of the five per-fold values. Bold values indicate the best results.

Method	Fold	AP ₅₀	AP ₇₅	mAP _{50:95}	F1 ₅₀	All-GT Dice	AJI+
Mask R-CNN	0	0.770	0.623	0.532	0.772	0.705	0.683
	1	0.761	0.636	0.531	0.767	0.683	0.665
	2	0.773	0.630	0.541	0.783	0.687	0.686
	3	0.786	0.622	0.511	0.768	0.667	0.673
	4	0.725	0.614	0.575	0.805	0.688	0.698
	Pooled	0.763	0.625	0.538	0.779	0.686	0.681
	Mean $\pm$ SD	0.763 $\pm$ 0.023	0.625 $\pm$ 0.008	0.538 $\pm$ 0.023	0.779 $\pm$ 0.016	0.686 $\pm$ 0.014	0.681 $\pm$ 0.013
SOLOv2	0	0.775	0.634	0.551	0.786	0.716	0.737
	1	0.812	0.644	0.543	0.789	0.707	0.712
	2	0.784	0.629	0.539	0.808	0.703	0.720
	3	0.768	0.634	0.537	0.799	0.720	0.729
	4	0.781	0.674	0.535	0.793	0.714	0.722
	Pooled	0.784	0.643	0.541	0.795	0.712	0.724
	Mean $\pm$ SD	0.784 $\pm$ 0.017	0.643 $\pm$ 0.018	0.541 $\pm$ 0.006	0.795 $\pm$ 0.009	0.712 $\pm$ 0.007	0.724 $\pm$ 0.009
Mask2Former	0	0.783	0.660	0.563	0.835	0.740	0.737
	1	0.787	0.650	0.580	0.826	0.748	0.757
	2	0.773	0.658	0.557	0.821	0.762	0.749
	3	0.772	0.654	0.580	0.826	0.736	0.737
	4	0.835	0.633	0.510	0.817	0.729	0.740
	Pooled	0.790	0.651	0.558	0.825	0.743	0.744
	Mean $\pm$ SD	0.790 $\pm$ 0.026	0.651 $\pm$ 0.011	0.558 $\pm$ 0.029	0.825 $\pm$ 0.007	0.743 $\pm$ 0.013	0.744 $\pm$ 0.009
MC-SlotNet	0	0.801	0.639	0.545	0.847	0.736	0.758
	1	0.815	0.651	0.559	0.865	0.761	0.754
	2	0.789	0.651	0.560	0.844	0.762	0.738
	3	0.795	0.649	0.554	0.845	0.793	0.770
	4	0.800	0.655	0.607	0.809	0.753	0.750
	Pooled	0.800	0.649	0.565	0.842	0.761	0.754
	Mean $\pm$ SD	0.800 $\pm$ 0.010	0.649 $\pm$ 0.006	0.565 $\pm$ 0.024	0.842 $\pm$ 0.020	0.761 $\pm$ 0.021	0.754 $\pm$ 0.012

3.2. Performance by Magnification

Table 7 reports MC-SlotNet’s principal metrics separately for each acquisition magnification, computed from pooled out-of-fold predictions restricted to each magnification subset. Performance improved monotonically from 4 $\times$ to 40 $\times$: AP₅₀ rose from 0.765 at 4 $\times$ to 0.860 at 40 $\times$, and All-GT Dice rose from 0.724 to 0.819 over the same range. This pattern is consistent with the higher per-image instance density and smaller apparent cell size at low magnification (Table 2) and supports the qualitative observations in Figure 5.

Table 7. MC-SlotNet performance by acquisition magnification (pooled out-of-fold predictions restricted to each magnification subset).

Magnification	AP ₅₀	AP ₇₅	mAP _{50:95}	F1 ₅₀	All-GT Dice	AJI+
4 $\times$	0.765	0.607	0.526	0.807	0.724	0.719
10 $\times$	0.805	0.657	0.571	0.852	0.769	0.761
20 $\times$	0.835	0.692	0.606	0.877	0.799	0.791
40 $\times$	0.860	0.717	0.631	0.895	0.819	0.811
Overall (pooled)	0.800	0.649	0.565	0.842	0.761	0.754

3.3. Overlap Recovery and Computational Efficiency

Overlap-specific evaluation further favored MC-SlotNet (Table 8). It achieved the highest all-ground-truth overlap Dice (0.705) and overlap-instance recall (0.828), together with the lowest FNR_0 at IoU 0.70 (0.251). Mask2Former achieved an overlap Dice of 0.695 and recall of 0.796, while SOLOv2 reached 0.683 and 0.786, respectively.

Table 8. Per-fold overlap recovery and full-image inference efficiency. “Pooled” is computed from all out-of-fold predictions combined; “Mean $\pm$ SD” is the unweighted average $\pm$ standard deviation across the five folds. Higher is better for overlap Dice and recall; lower is better for FNR_0 and runtime. Bold values indicate the best results.

Method	Fold	Overlap Dice	Overlap Recall	$\text{FNR}_{0.70}$	Runtime (s/Image)
Mask R-CNN	0	0.702	0.751	0.398	6.227
	1	0.664	0.751	0.408	6.328
	2	0.673	0.716	0.423	6.306
	3	0.679	0.771	0.406	6.314
	4	0.642	0.766	0.425	6.325
	Pooled	0.672	0.751	0.412	6.300
	Mean $\pm$ SD	0.672 $\pm$ 0.022	0.751 $\pm$ 0.022	0.412 $\pm$ 0.012	6.300 $\pm$ 0.042
SOLOv2	0	0.713	0.757	0.285	1.286
	1	0.686	0.778	0.325	1.259
	2	0.676	0.740	0.265	1.392
	3	0.720	0.807	0.313	1.407
	4	0.620	0.848	0.317	1.251
	Pooled	0.683	0.786	0.301	1.319
	Mean $\pm$ SD	0.683 $\pm$ 0.040	0.786 $\pm$ 0.043	0.301 $\pm$ 0.025	1.319 $\pm$ 0.075
Mask2Former	0	0.710	0.760	0.363	3.427
	1	0.692	0.804	0.326	3.370
	2	0.705	0.841	0.323	3.467
	3	0.681	0.801	0.327	3.437
	4	0.687	0.774	0.266	3.509
	Pooled	0.695	0.796	0.321	3.442
	Mean $\pm$ SD	0.695 $\pm$ 0.012	0.796 $\pm$ 0.031	0.321 $\pm$ 0.035	3.442 $\pm$ 0.051
MC-SlotNet	0	0.698	0.810	0.284	1.102
	1	0.671	0.804	0.245	1.130
	2	0.703	0.807	0.253	1.150
	3	0.736	0.822	0.283	1.103
	4	0.717	0.897	0.190	1.080
	Pooled	0.705	0.828	0.251	1.113
	Mean $\pm$ SD	0.705 $\pm$ 0.024	0.828 $\pm$ 0.039	0.251 $\pm$ 0.038	1.113 $\pm$ 0.027

Runtime was measured on a single NVIDIA A40 GPU. The timed interval included physical-scale resampling, normalization, tiled inference, mask decoding, tile merging, duplicate suppression, and rescaling to the native image resolution. Reported runtime variability reflects differences in image workload, particularly the number of tiles required after physical-scale normalization. MC-SlotNet was also the fastest evaluated model, requiring 1.113 s per full image compared with 1.319 s for SOLOv2, 3.442 s for Mask2Former, and 6.300 s for Mask R-CNN. Thus, its improved overlap recovery was not obtained at the expense of increased inference time.

3.4. Qualitative Full-Cell Segmentation

Figure 5 compares representative predictions at 40 $\times$, 20 $\times$, 10 $\times$, and 4 $\times$ magnification. Across the four scales, all methods recovered many clearly separated cells, whereas differences became more apparent in dense clusters, overlapping regions, and cells intersecting image boundaries.

Across magnifications, all methods performed well on clearly separated cells, while larger differences emerged in crowded regions, overlap zones, and image boundaries. MC-SlotNet generally preserved more complete full-cell outlines and recovered more overlapping instances, particularly at 20 $\times$ and 10 $\times$. Mask2Former remained competitive

but showed greater boundary disagreement in some overlapping groups, while SOLOv2 and Mask R-CNN more often produced incomplete or missed instances in dense regions.

At $4\times$, where apparent cell size was smallest and image density highest, all methods became less reliable. Nevertheless, MC-SlotNet retained a dense set of individual masks and preserved many overlapping instances. These observations are consistent with the quantitative overlap-recovery results and should be interpreted as qualitative examples rather than independent evidence of performance.

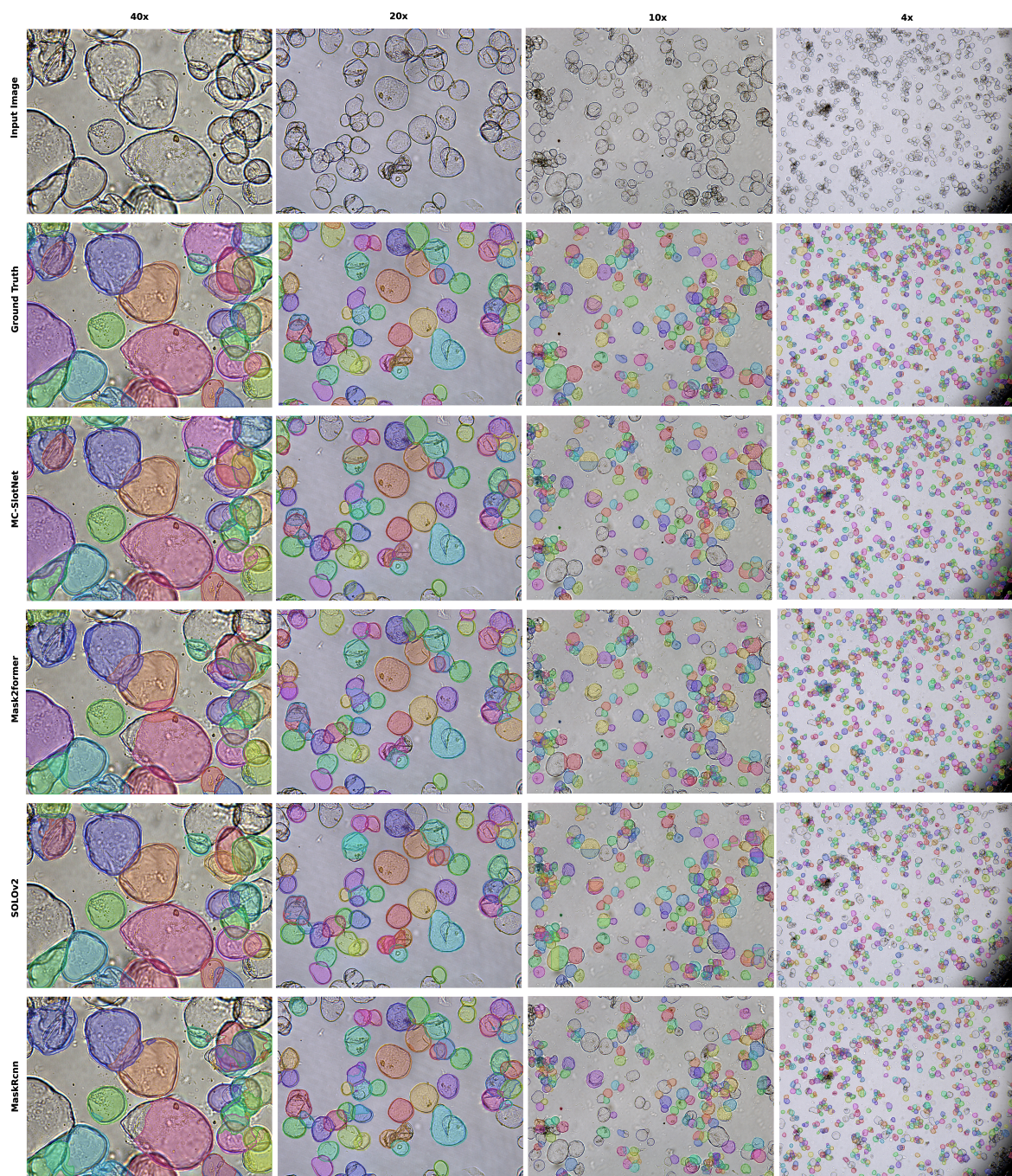


Figure 5. Qualitative comparison of full-cell instance segmentation at $40\times$, $20\times$, $10\times$, and $4\times$ magnification. Columns correspond to magnification and rows show the microscopy inputs, independently annotated ground-truth full-cell masks, and predictions from MC-SlotNet, Mask2Former, Mask R-CNN, and SOLOv2. Colored overlays represent independent instance masks, allowing shared pixels to belong to more than one predicted cell. The examples illustrate differences in full-cell shape recovery, crowded-region segmentation, and border-instance detection across imaging scales.

3.5. Competitive Attention and Full-Cell Reconstruction

Figure 6 illustrates how competitive feature assignment is separated from final spatial occupancy in MC-SlotNet. The ground-truth multiplicity map highlights pixels shared by multiple annotated cells. For Slots 0, 6, and 8, the competitive attention maps remain relatively broad and include contextual responses from neighboring structures, whereas the independently decoded full-cell probabilities are sharply localized to the corresponding cells.

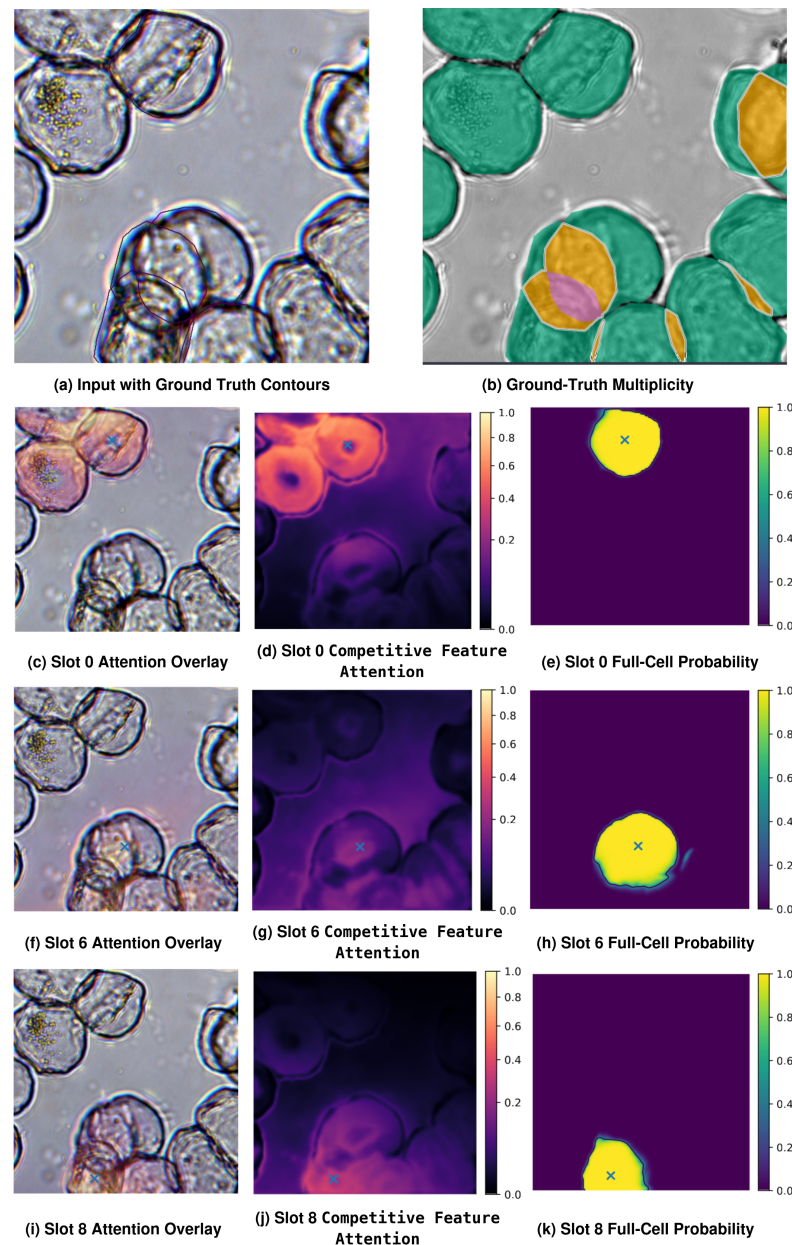


Figure 6. Qualitative visualization of competitive feature attention and independent full-cell reconstruction for a representative held-out crop. (a) Microscopy image with ground-truth full-cell contours and (b) ground-truth cell multiplicity. Rows corresponding to Slots 0, 6, and 8 show the slot-specific attention overlay, competitive feature-attention map, and independently decoded full-cell probability. Crosses indicate the associated query centers. Attention represents competitive feature assignment during slot refinement and should not be interpreted as an instance-segmentation probability. The independently decoded masks represent final cell occupancy and may overlap spatially. For visualization only, all attention maps were displayed using the same power-law normalization ($\gamma = 0.75$, range 0–1). Blue cross marks represent cell center.

This behavior is consistent with the intended architecture: attention determines how spatial features contribute to slot refinement but is not itself an instance mask. Final occupancy is instead represented by independent dynamic masks, allowing masks from different slots to overlap when required by the full-cell annotations.

3.6. Ablation Experiments

The contribution of competitive feature assignment and mask-level multiplicity consistency was evaluated using controlled ablations in which the backbone, data split, training schedule, and inference procedure were held fixed. The results of these experiments are reported in Table 9.

Table 9. Ablation of attention competition and mask-level multiplicity consistency in MC-SlotNet. A4 additionally uses multiplicity-guided sigmoid attention and attention-budget supervision.

Variant	Attention Formulation	Mask Mult.	AP ₅₀	AP ₇₅	mAP _{50:95}	AJI+	Overlap Dice	Overlap Recall
A0	Softmax	–	0.792	0.642	0.557	0.744	0.690	0.807
A1 (proposed)	Softmax	✓	0.800	0.649	0.565	0.754	0.705	0.828
A2	Sigmoid	–	0.787	0.635	0.550	0.739	0.687	0.812
A3	Sigmoid	✓	0.795	0.642	0.558	0.748	0.699	0.823
A4	Sigmoid + multiplicity-guided attention	✓	0.789	0.636	0.552	0.743	0.696	0.819

Relative to the competitive baseline A0, the proposed A1 formulation improved mAP_{50:95} from 0.557 to 0.565 and produced larger gains in overlap Dice (0.690 to 0.705) and overlap recall (0.807 to 0.828), suggesting that mask-level multiplicity consistency was particularly beneficial for reconstruction of shared cell regions. Replacing competitive softmax attention with independent sigmoid attention reduced overall segmentation performance in A2 and A3, although mask consistency again improved overlap recovery when comparing A2 with A3. A4 achieved AP₅₀ = 0.789, compared with 0.795 for A3. Both variants used sigmoid attention and mask-level multiplicity consistency, but A4 additionally used the predicted multiplicity to guide attention and supervised the aggregate attention occupancy. The absence of an improvement suggests that explicitly imposing multiplicity at the attention stage was unnecessary once the final mask set was already constrained by mask-level multiplicity consistency. One possible explanation is that encouraging multiple slots to respond strongly to the same high-multiplicity regions can reduce object-specific slot specialization. These results therefore support the simpler design adopted in A1, in which attention is used for object-feature assignment while multiplicity supervision is applied to the final independent mask reconstruction.

4. Discussion

The results showed the importance of representing overlapping cells as independent full-cell instances rather than forcing a single label at each pixel. In this study, MC-SlotNet achieved the highest numerical mAP_{50:95}, F1, all-ground-truth Dice, AJI+, overlap Dice, and overlap recall, while requiring the lowest full-image inference time. Its advantage was therefore observed both in general instance recovery and specifically in regions where annotated cells shared pixels.

Mask2Former remained the closest competing method. Its AP₇₅ of 0.651 slightly exceeded the 0.649 obtained by MC-SlotNet, indicating that the two models were particularly close under strict localization. This result is important because Mask2Former provides a strong pretrained query-based instance-segmentation reference rather than an artificially weak baseline. SOLOv2 also performed competitively and provided the second-fastest

inference, supporting its inclusion as a more relevant one-stage baseline than an exclusive watershed pipeline.

The small AP_{75} gap (0.651 for Mask2Former versus 0.649 for MC-SlotNet) is consistent with a difference in how the two architectures localize object boundaries. Mask2Former's masked cross-attention restricts each query's attention to its own predicted mask region at every decoder layer, which tends to sharpen boundaries under strict localization. MC-SlotNet instead decodes each full-cell mask independently from a shared mask-feature map using a lightweight per-slot linear/quadratic readout (Equation (9)), which favors consistent recovery of overlapping instances but can slightly over-smooth the shared boundary between two heavily overlapping cells. This is visible in the $40\times$ panels of Figure 5, where MC-SlotNet recovers more of the overlapping instances overall but occasionally traces a marginally softer shared boundary than Mask2Former.

The qualitative comparison supported the quantitative trends across the full $4\times$ – $40\times$ acquisition range. MC-SlotNet generally preserved coherent full-cell outlines in both isolated and overlapping regions, with particularly visible advantages in dense clusters and several border-adjacent instances. At lower magnifications, where individual cells occupied fewer pixels and cell density increased substantially, all methods showed more omissions and boundary deviations. Nevertheless, MC-SlotNet continued to recover a dense set of independent instances at $10\times$ and $4\times$, including masks participating in substantial overlap. These observations suggest that the combination of independent mask decoding and multiplicity-aware supervision remains useful as apparent cell size and local crowding change across magnifications.

The attention visualization further clarifies that competitive attention is not itself the final segmentation. Attention remains spatially contextual during slot refinement, whereas the subsequent dynamic mask decoder produces an object-specific full-cell occupancy estimate. This separation allows slots to compete for feature representation without preventing their independently decoded masks from sharing pixels. Physical-scale normalization enabled a common processing scale across $4\times$ – $40\times$ images, but it faces difficulty recovering information absent at lower magnification.

Error analysis showed that failures were concentrated in difficult structural conditions rather than in a single magnification. MC-SlotNet obtained $FNR_o = 0.251$ at IoU 0.70, while the overlap-instance recall of 0.828 indicates that 17.2% of annotated cells participating in overlap were not recovered at IoU 0.50. Qualitative inspection identified five recurring challenges: very small apparent cells at lower magnification, weak or refractive boundaries, extreme local crowding, cells intersecting image borders, and occasional incomplete reconstruction of strongly overlapping outlines. These observations are descriptive and are not used as substitutes for the quantitative benchmark.

Several limitations constrain the scope of the present findings. The dataset contains only 53 annotated images from one plant species and one brightfield acquisition system, with a strongly imbalanced distribution across magnifications. Morphologically different plant or animal cells and fluorescence, phase-contrast, or confocal modalities may therefore require additional training or adaptation. Physical-scale normalization enables a common processing scale but cannot restore spatial information that is absent at low magnification and may suppress fine high-resolution detail. In addition, full-cell boundaries in densely overlapping regions were reconstructed by a domain expert and cross-checked by a second reviewer rather than through a formal multi-annotator consensus protocol, and inter-annotator agreement was not quantitatively assessed. Some subjectivity in tracing ambiguous or heavily occluded cell boundaries is therefore possible and should be considered when interpreting absolute segmentation accuracy. Finally, the present formulation is two-dimensional: it reconstructs annotated projected full-cell outlines but does

not infer front-to-back depth ordering, hidden three-dimensional surfaces, or true cell volume. Extension to z-stacks would require volumetric annotations and a three-dimensional instance model.

5. Conclusions

We introduced MC-SlotNet, an overlap-aware instance-segmentation framework that combines competitive object-slot refinement with independent full-cell mask decoding and mask-level multiplicity supervision. The framework was evaluated on a newly annotated suspension-culture microscopy dataset comprising 53 micrographs and 4131 full-cell instances with overlap-preserving annotations. Across grouped five-fold cross-validation, MC-SlotNet outperformed the evaluated baselines on most instance- and overlap-specific metrics while achieving the lowest full-image inference time. Mask2Former achieved a marginally higher AP₇₅ (0.651 versus 0.649), whereas MC-SlotNet showed stronger overall performance across the remaining evaluation criteria.

These results suggest that explicitly representing overlapping cells as independent full-cell instances is well suited to suspension-culture microscopy, where projected overlap violates the usual assumption of mutually exclusive object masks. The findings further show that overlap-preserving annotation and evaluation can reveal performance differences that may be obscured by conventional instance-segmentation protocols.

Author Contributions: Conceptualization, T.U.R., S.L. and M.R.U.R.; methodology, T.U.R., S.L. and M.R.U.R.; software, T.U.R., S.L., M.T.S. and M.R.U.R.; validation, M.G. and M.R.U.R.; formal analysis, T.U.R. and S.L.; investigation, T.U.R. and S.L.; resources, M.G. and M.R.U.R.; data curation, T.U.R.; writing—original draft preparation, T.U.R.; writing—review and editing, T.U.R., S.L., M.T.S., M.G. and M.R.U.R.; visualization, S.L. and M.T.S.; supervision, M.G. and M.R.U.R.; project administration, M.G. and M.R.U.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset presented in this article is not readily available because the annotated microscopy data form part of an ongoing study and are currently being further curated and extended. The training and inference code for this study is likewise not publicly released at this stage, as it is being actively developed for follow-on work on this dataset. Requests to access the data or code should be directed to the corresponding author.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.6 Sol; OpenAI), to improve the language, grammar, clarity, and readability of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kalia, A. Chapter 12-Nanotechnology in Bioengineering: Transmogrifying Plant Biotechnology. In *Omics Technologies and Bio-Engineering*; Barh, D., Azevedo, V., Eds.; Academic Press: New York, NY, USA, 2018; pp. 211–229. [\[CrossRef\]](#)
2. Bapat, V.A.; Kavi Kishor, P.B.; Jalaja, N.; Jain, S.M.; Penna, S. Plant Cell Cultures: Biofactories for the Production of Bioactive Compounds. *Agronomy* **2023**, *13*, 858. [\[CrossRef\]](#)
3. Ibaraki, Y.; Kenji, K. Application of image analysis to plant cell suspension cultures. *Comput. Electron. Agric.* **2001**, *30*, 193–203. [\[CrossRef\]](#)
4. Zhu, X.; Mohsin, A.; Zaman, W.Q.; Liu, Z.; Wang, Z.; Yu, Z.; Tian, X.; Zhuang, Y.; Guo, M.; Chu, J. Development of a novel noninvasive quantitative method to monitor *Siraitia grosvenorii* cell growth and browning degree using an integrated computer-aided vision technology and machine learning. *Biotechnol. Bioeng.* **2021**, *118*, 4092–4104. [\[CrossRef\]](#) [\[PubMed\]](#)

5. Ercetin, A.; Arpacı, I.; Memiş, S.; Akagıç, A.; Özgün, Ö. AI-Driven Image Processing for Microstructure and Surface Characterization: A Systematic Review of Methods, Materials, and Applications. *Arch. Comput. Methods Eng.* **2026**, 1–28. [\[CrossRef\]](#)
6. Hafız, A.M.; Bhat, G. A survey on instance segmentation: State of the art. *Int. J. Multimed. Inf. Retr.* **2020**, *9*, 171–189. [\[CrossRef\]](#)
7. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Munich, Germany, 5–9 October 2015; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9351, pp. 234–241. [\[CrossRef\]](#)
8. Schmidt, U.; Weigert, M.; Broaddus, C.; Myers, G. Cell Detection with Star-Convex Polygons. In Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2018, Granada, Spain, 16–20 September 2018; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2018; Volume 11071, pp. 265–273. [\[CrossRef\]](#)
9. Stringer, C.; Wang, T.; Michaelos, M.; Pachitariu, M. Cellpose: A Generalist Algorithm for Cellular Segmentation. *Nat. Methods* **2021**, *18*, 100–106. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Wolny, A.; Cerrone, L.; Vijayan, A.; Tofanelli, R.; Barro, A.V.; Louveaux, M.; Wenzl, C.; Strauss, S.; Wilson-Sánchez, D.; Lymbouridou, R.; et al. Accurate and Versatile 3D Segmentation of Plant Tissues at Cellular Resolution. *eLife* **2020**, *9*, e57613. [\[CrossRef\]](#) [\[PubMed\]](#)
11. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969. [\[CrossRef\]](#)
12. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. Solov2: Dynamic and fast instance segmentation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17721–17732.
13. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 1280–1289. [\[CrossRef\]](#)
14. Li, L.; Chen, W.; Qi, J. VB-SOLO: Single-Stage Instance Segmentation of Overlapping Epithelial Cells. *IEEE Access* **2024**, *12*, 52555–52564. [\[CrossRef\]](#)
15. Prytula, Y.; Tsioporenko, I.; Zeynalli, A.; Fishman, D. IAUNet: Instance-Aware U-Net. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Nashville, TN, USA, 11–15 June 2025; pp. 4778–4787.
16. Kirkegaard, J.B. Spontaneous Breaking of Symmetry in Overlapping Cell Instance Segmentation Using Diffusion Models. *Biol. Methods Protoc.* **2024**, *9*, bpae084. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Li, K.; Malik, J. Amodal Instance Segmentation. In Proceedings of the Computer Vision–ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2016; Volume 9912, pp. 677–693. [\[CrossRef\]](#)
18. Rehman, T.U.; Rahman, M.R.U.; Tang, W.; Vascon, S.; Jiang, P.; Liu, Y.; Gong, S.; Wan, X.; Mohsin, A.; Guo, M. ExoOrb: A novel visual and analytical system for therapeutic extracellular vesicles metrics. *Comput. Struct. Biotechnol. J.* **2025**, *27*, 5289–5306. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A Database and Web-Based Tool for Image Annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173. [\[CrossRef\]](#)
20. 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 Proceedings of the Computer Vision–ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693, pp. 740–755. [\[CrossRef\]](#)
21. Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.Y.; Cubuk, E.D.; Le, Q.V.; Zoph, B. Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2918–2928. [\[CrossRef\]](#)
22. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022. [\[CrossRef\]](#)
23. Locatello, F.; Weissenborn, D.; Unterthiner, T.; Mahendran, A.; Heigold, G.; Uszkoreit, J.; Dosovitskiy, A.; Kipf, T. Object-Centric Learning with Slot Attention. In Proceedings of the Advances in Neural Information Processing Systems, Virtually, 6–12 December 2020; Volume 33, pp. 11525–11538.
24. Tian, Z.; Shen, C.; Chen, H. Conditional Convolutions for Instance Segmentation. In Proceedings of the Computer Vision–ECCV 2020, Virtually, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 282–298. [\[CrossRef\]](#)
25. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850. [\[CrossRef\]](#)
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.

27. Kumar, N.; Verma, R.; Sharma, S.; Bhargava, S.; Vahadane, A.; Sethi, A. A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. *IEEE Trans. Med. Imaging* **2017**, *36*, 1550–1560. [[CrossRef](#)] [[PubMed](#)]
28. Graham, S.; Vu, Q.D.; Raza, S.E.A.; Azam, A.; Tsang, Y.W.; Kwak, J.T.; Rajpoot, N. HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images. *Med. Image Anal.* **2019**, *58*, 101563. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.