Reviewer 1:

Summary:

In this paper, the authors introduce a method for detecting cell types in multiplex immunofluorescence (mIF) imaging. The approach first involves training a model, specifically YOLOv8, using a limited set of annotations. This trained model is then used to generate pseudo-labels, which are in turn used to retrain the model. The authors have conducted extensive experiments across various images representing different types of cancers and tissues.

Strengths:

- The authors have conducted experiments across a diverse range of cancer subtypes, accounting for tissue heterogeneity across different organs and cancer types.
- The paper addresses a notably challenging task of cell detection in multiple immunofluorescence (mIF) imaging, particularly within the constraints of limited annotations.

Weaknesses:

- The paper lacks a comparative analysis with state-of-the-art (SOTA) methods. For instance, the method could be compared with models such as Stardist, or Cellpose, pre-trained on immunofluorescence (iF) data and fine-tuned with the annotations provided here, with a simple thresholding method for cell classification.

We thank the reviewer for the comment. Since the objective of this work was cell detection and classification, we focused on comparing against models developed for detection, such as Fast R-CNN, YOLOv5, and YOLOv8. Cellpos and Stardist are cell segmentation models and not directly applicable here. Stardist expects full cellular masks and it doesn't match our annotations. Without finetuning, the trained nuleus detection Stardist model applied to the blue channel in our mIF images (makers for nucleus), obtains an F1-Score of 0.35 in the validation dataset. We associate this poor performance with characteristics of mIF imaging such as poor definition of cell boundaries as well as small sizes of cells in our data. For finetuning starDist, we adjusted our labels by expanding the centroids of annotations by 50 pixels and creating circular masks for each annotation and achieved the average precision and recall scores of 0.284 across cell types. Given this poor performance compared to Yolov8 (Table 2, 0.848), we decided to not include these results in the manuscript as we believe this quick data preparation approach may have not been sufficient to provide StarDist a fair chance.

- The performance of YOLOv8 trained solely on partial annotations is not discussed for all five cancer types in the evaluation dataset, which limits the ability to directly assess the interest of the proposed method compared to using the model trained on limited annotated data alone, which appears to already exhibit satisfactory performance.

Thank you for your feedback. As suggested by the reviewer, to evaluate the contribution of pseudo labels to improve the task performance across the five cancer types, in this version we compare the performance of models with and without pseudo labels. We show that incorporating pseudo labels generation significantly improves F1 score (CD45: 0.656 vs. 0.949, panCK: 0.417 vs. 0.880, Others: 0.359 vs.0.740) with p values (Wilcoxon test) of 3.0517578125e-05, 1.9073486328125e-06, 3.0517578125e-05, respectively. We have incorporated these results in Figure 3A and section 3.2

- The description of the experimental setup is missing critical details, such as the total number of pseudo-labels generated, the methodology used to tune this parameter, and the assessment of pseudo-label quality. The paper mentions strategies like Consistency-based Semi-supervised learning for object Detection (CSD) to prevent overfitting but fails to clarify if they were implemented or how they were adapted. The potential for overfitting is a significant concern with methods of this nature, and the results indicate some degree of overfitting, particularly in the "others" cell type across the four additional cancer types, raising important questions about the method's robustness.

Thank you for this feedback. We have clarified the increase in dataset size after incorporating pseudo-labels: from initial counts of 11,643 for CD45, 15,228 for panCK, and 9,489 for Others, to 140,000, 200,000, and 120,000 for CD45, panCK, and Others, respectively. Also, we have updated Section 2.3 to address concerns about parameter tuning, pseudo-label quality, and overfitting. This procedure includes our method for creating pseudo-labels. We guaranteed the accuracy of these pseudo-labels by only accepting those with high confidence, that is, predictions with a confidence level of over 90%. This means new predicted labels were added only if their prediction confidence threshold was above 90%. As for mitigating overfitting we adopted strategies including dropout, data augmentation, regularization, and early stopping. Furthermore, the model's performance was evaluated across five fully annotated cancer types not encountered during the training phase, resulting in average precision, recall, and F1 scores of 0.904, 0.851, and 0.848, respectively, for cell detection (calculated based on table 2 in manuscript). Also, the evaluation was conducted on a limited sample of four patches per cancer type. Considering the heterogeneity of tumors, this sample size may not have been sufficient to capture the full spectrum of variability present in each cancer type, potentially leading to an imbalanced testing set. This imbalance might have contributed to the lower performance observed in the 'Others' category for Urothelial Carcinoma and cholangiocarcinoma.

- The clarity and structure of the paper could be improved, as it is currently challenging to follow. The blending of experimental settings with methods and results, coupled with a literature review that is not organized chronologically, hinders the reader's comprehension of the paper, and makes it difficult to clearly identify the contribution of the authors.

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct
Detailed Comments:

N/A

Questions To Address In The Rebuttal:

To qualify for publication at MIDL2024, the authors must improve the overall clarity and structure of the manuscript, clearly highlighting the novel contributions of their work. Additionally, it is imperative to undertake further experiments that compare their approach with state-of-the-art (SOTA) methods and evaluate the effectiveness of the model when trained solely on partial annotations on the evaluation datasets. These steps are essential to demonstrate the value and efficacy of the proposed semi-supervised learning method.

Preliminary Rating: 1: Strong reject
Justification Of the Preliminary Rating:

At this current stage, the paper does not meet the required criteria for publication at MIDL2024, and the number of issues that urgently need to be addressed appears far too extensive for the allotted rebuttal period.

Reviewer 2:

Summary:

This paper discusses a semi-supervised approach developed to improve cell detection in mIF within the tumor microenvironment. The challenge arises from limited and unevenly distributed annotations for training cell detectors. The authors tested three object detection models with tremendous partially annotated data from different cancer types. An enriched dataset was created using pseudo labels generated by YOLOv8s. The fine-tuned model achieved high accuracy on fully annotated data from five cancer types, demonstrating the effectiveness of the semi-supervised approach in cellular analysis of mIF.

Strengths:

- The annotation is incredibly large.
- The paper covers various diseases with clinical relevance.
- The model selection is easy to follow and should aid in the implementation of other potential marker combinations.

Weaknesses:

- The representation could benefit from some improvement. For instance, the description accompanying Figure 1 may require enhanced clarity to better convey the process of training all detectors on partially annotated datasets and the subsequent selection of the separate top-performing model for pseudo label generation through an iterative loop. The paper repeatedly claims that the improvement is significant. However, there is no statistical significance test provided as evidence.

Confidence: 5: The reviewer is certain that the evaluation is correct and very familiar with the relevant literature
Detailed Comments:

- Inclusion of the Area Under the Curve (AUC) performance metrics in the appendix is recommended for a more comprehensive evaluation of the model's discriminative ability

Thank you for your valuable feedback. As suggested, in this version we have incorporated the AUC metrics (0.90, 0.80, 0.83 for CD45, panCK, and Others, respectively) into Appendix C and referenced it in Section 3.2 of the results for a more thorough evaluation of the model's discriminative capacity.

- Concerning the generation of pseudo labels and the potential for cascading errors, it would be prudent for the authors to outline the measures implemented to mitigate error propagation from the pseudo label generator to the final model. Rigorous cross-validation, confidence thresholding, or incorporating expert review in iterations could be potential strategies to ensure the reliability and accuracy of the data used for further training.

Thank you for your great feedback. In this version we have clarified the details related to parameter tuning, pseudo-label quality, and overfitting. We ensured reliability of pseudo labels by including only high-confidence predictions (outputs with > 90%). To evaluate if errors during pseudo labeling generation would propagate to the model's performance, we compare the performance of two approaches, with and without pseudo labels, in classifying cells. In this version (Figure 3A) we show that in data unseen during training, the addition of pseudo labels boosted the model performance; hence, resulting very unlikely a propagation of error during the pseudo label generation step that would have a significant impact of cell detection and classification.

- Figure 1 would benefit from adding the unit of measurement, specifically micrometers, to clarify the scale of observation.

Thanks for feedback, unit is added as requested.

Questions To Address In The Rebuttal:

Please review the weaknesses and detailed comments sections. Cross-validation is encouraged but not mandatory during the rebuttal phase.

Preliminary Rating: 3: Borderline
Justification Of The Preliminary Rating:

The need for improved clarity in figures and methodological rigor, including statistical validation of significant claims.

The importance of addressing the propagation of errors from pseudo labels to the final model .

I also suggest providing additional quantitative metrics, such as AUC performance, to substantiate the paper's findings.

Thank you for your feedback. As suggested by the reviewer, to evaluate the contribution of pseudo labels to improve the task performance across the five cancer types, in this version we compare the performance of models with and without pseudo labels. We show that incorporating pseudo labels generation significantly improves F1 score (CD45: 0.656 vs. 0.949, panCK: 0.417 vs. 0.880, Others: 0.359 vs.0.740) with p values (Wilcoxon test) of 3.0517578125e-05, 1.9073486328125e-06, 3.0517578125e-05, respectively. We have incorporated these results in Figure 3A and section 3.2, Also, we have added the Area Under the Curve (AUC) performance metrics into Appendix C and referenced it in Section 3.2 of the results for a more thorough evaluation of the model's discriminative capacity.

Reviewer 3:

Summary:

In this paper, the authors propose to enhance cell detection in low-data regimes for multiplex immunofluorescence imaging by using self-distillation. They train three different detection models on a small manually annotated dataset and chose YOLOv8s of these as the best-performing one. In the next step, they predict unseen data by the model to generate pseudo-labels. Another YOLOv8s model is trained using the combination of manual and pseudo-labels. The authors evaluate the resulting model on fully, manually annotated test data of 5 different cancer types.

Strengths:

- Utilizing little manually annotated training data is a relevant topic to all medical imaging problems.
- The authors highlight that the test data is taken from different samples than training data.
- Testing is based on 5 different cancer types of which 4 have not been included in training.

Weaknesses:

- the paper is not very well structured. Examples are.

- in the introduction, related work for cell segmentation is discussed (Greenwald et al), then cell classification (Amitay et al. etc) and then segmentation again (Schmidt et al. etc)

Thank you for your valuable feedback. We have carefully restructured the introduction section to ensure that all related literature is presented cohesively and consistently.

- The results section contains a lot of descriptions on the evaluation process, which would belong to an evaluation section and a lot of result interpretation, which would belong to the discussion section.

- There are quite a few redundancies in the paper.
- The experiment of the different annotation levels is entirely in the appendix, still its results are referred to in the discussion.

Thank you for your valuable feedback, based on your suggestion, we have now removed these references from the discussion.

- In the whole paper, it is not clear if the model only detected cells or classified (line 123). Are there 3 different classes the model must decide between or is just the evaluation separated into those classes?

Thanks for your comments, our model is designed for both detection and classification of cells. As illustrated in Figure 1, the model differentiates among three categories: immune cells (CD45+), epithelial and cancer cells (panCK+), and others (CD45-panCK-). This clarifies that the evaluation encompasses these specific classes, detailing our method's capacity to not only detect but also accurately classify cells into these predetermined categories.

- The results of the proposed approach are not compared to any other approaches. Comparison to state-of-the-art methods or comparison to the YOLOv8s model without training on the pseudo-labels would have been useful.

Thank you for your feedback. As suggested by the reviewer, to evaluate the contribution of pseudo labels to improve the task performance across the five cancer types, in this version we compare the performance of models with and without pseudo labels. We show that incorporating pseudo labels generation significantly improves F1 score (CD45: 0.656 vs. 0.949, panCK: 0.417 vs. 0.880, Others: 0.359 vs.0.740) with p values (Wilcoxon test) of 3.0517578125e-05, 1.9073486328125e-06, 3.0517578125e-05, respectively. We have incorporated these results in Figure 3A and section 3.2, Also, we have added the Area Under the Curve (AUC) performance metrics into Appendix C and referenced it in Section 3.2 of the results for a more thorough evaluation of the model's discriminative capacity.

- The paper lacks novelty, as the methods are already state of the art (as the authors explain in the introduction). Also, the validation is very limited such that there are only limited new findings on the application of the method on mIF images (will be clarified).

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Detailed Comments:

- The references to Hinton et al, is at the wrong position, it should be after the term "Self-distillation" and not after its application to immune infiltration.

Thanks, the reference location has been modified

- Evaluation metrics are not explained. mAP50 should at least have one sentence of explanation or a reference to one.

Thank you. We have added this part (mean Average Precision at 50% Intersection Over Union (IoU)) to the sentence: “The table presents key measures of model performance, recall, and mAP50 (mean Average Precision at 50% Intersection Over Union (IoU)), which are central to our evaluation criteria.”

- The manual annotations should be described in some detail. Is it point or bounding box annotations? How do you annotate 10% of the data? Is it 10% of the images or 10% of the cells? If 10% of the cells, how can false positives be measured to calculate precision?

Thank you for your valuable feedback. To clarify the annotation and model training process further, we have updated the methods section of our manuscript as follows: After pathologists annotated the center of cells by marking approximately 10% of the cells across 10 images, we proceeded to draw 50*50 pixels bounding box around each point. In total, pathologists annotated 11,643 cells as CD45+, 15,228 as panCK+, and 9,489 as Others. This step was crucial for adapting the point annotations for training object detection models, which require bounding box inputs. For the evaluation of our model, we compared the predicted bounding boxes with this ground truth bounding boxes to accurately measure false positives rate and calculate precision, ensuring the model's effectiveness in detecting and classifying cell types.

- "Furthermore, the study incorporated additional fully annotated datasets of five cancer types for final validation, enriching the robustness and adaptability of the models to different cancer cell appearances and histology conditions". This is very imprecise language. Performing some final validation by definition does not change the model and therefore cannot enrich robustness and adaptability of the models.

We have revised the sentence for clarity: "The study also incorporated fully annotated datasets from five distinct cancer types during the final validation phase. This was to evaluate the model's ability to perform across a range of cancer cell appearances and histological conditions."

- "We evaluated the generated pseudo labels through a designed loop by comparing them with the ground truth." is this just the description of how evaluation is performed? Otherwise, it is not clear what exactly is compared and what the result and the consequence of the comparison is.

Thanks for great feedback, we have updated Section 2.3 to address concerns about parameter tuning, pseudo-label quality, and overfitting. Our approach ensures reliable pseudo labels by including only high-confidence predictions (over 90%). The final model's performance across five distinct cancer types, which were fully annotated by expert pathologists not seen during training, attests to our method's robustness and effectiveness against overfitting.

- The authors state that YOLOv5s performs like YOLOv8s, while in the very next sentence they claim that the results highlight YOLOv8s' superiority. This is contradictory.

Thanks for feedback, kindly we revised to this sentence to avoid contradiction: "YOLOv5s also demonstrated high mAP50 and competitive recall scores, nearly matching those of YOLOv8s in certain aspects. However, the comprehensive analysis highlights YOLOv8s's slight edge in performance, especially in terms of recall, marking it as the most accurate and reliable among the tested models."

- The reference to Qu et al in the Conclusion is redundant and already done in the introduction.

Thanks for the feedback; we removed the repetitive reference.

Questions To Address In The Rebuttal:

The authors should be very clear on the contributions of their work and support their claims by performing the required evaluations. A rewrite of their paper following a clear structure would then help to get this message across

Preliminary Rating: 1: Strong reject
Justification Of The Preliminary Rating:

The contributions of the paper does not become clear. It does not present a novel method nor does it validate an existing method thoroughly. Additionally, the structure of the writing is often confusing, redundant, and lacking important details.