**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.

My Suggestion: explain about methods for reviewers which is in our study annotation is just for object detections and we do not have ground truth mask to comparing with these networks, also we compared with faster RCNN and YOLOv5 as well the results as well, also we used cellpose for some of our validation patches and got these results, our data in composite images and no single channel.

A close-up of a screenshot

Description automatically generated

* 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.

Figure 3 and line 208 to 212.

* The description of the experimental setup is missing critical details, such as the total number of pseudo-labels generated (line 164 to 168), the methodology used to tune this parameter (line 150 to 154), and the assessment of pseudo-label quality (line 159 to 163). 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 (line 150 to 159). 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 (line 150 to 159)
* 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. (will be improved)

**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 (line 114 to 119).
* 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 (will be improved).
* 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 (line 162 to 166).
* Figure 1 would benefit from adding the unit of measurement, specifically micrometers, to clarify the scale of observation (figure 1 modified).

**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 (figure 3).

The importance of addressing the propagation of errors from pseudo labels to the final model (line 162 to 166).

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

**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) (Modified by rearranging).
  + The results section contains a lot of description on the evaluation process, which would belong to an evaluation section and a lot of result interpretation, which would belong to the discussion section (modified).
* There is quite some redundancy in the paper (modified).
* The experiment of the different annotation levels is entirely in the appendix, still its results are referred to in the discussion (removed).
* In the whole paper it is not clear whether the model performed only detection of cells or also classification (line 123). Are there 3 different classes the model must decide between or is just the evaluation separated into those classes? (line 123).
* 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 (Figure 3 and line 208 to 212).
* 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. (modified).
* Evaluation metrics are not explained. mAP50 should at least have one sentence of explanation or a reference to one (line 192).
* 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? (lines 130 to 134 ).
* "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 (modified lines 176 to 179).
* "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 (lines 152 to 166).
* The authors state that YOLOv5s performs like YOLOv8s, while in the very next sentence they claim that the results highlight YOLOv8s' superiority. This is contradicting.
* The reference to Qu et al in the Conclusion is redundant and already done in the introduction (Modified lines 197 to 199).

**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.